Vol.2019-MUS-123 No.15
Vol.2019-SLP-127 No.15
2019/6/22

リアルタイム音声合成を用いたビブラートデザイン支援インタフェースの開発

小野 雄大[1,a)]　森勢 将雅[2]

概要：VOCALOID などの歌詞と譜面の情報から歌声を合成する歌声合成ソフトウェアを皮切りに，歌声合成技術は発展し続け，歌声に多様な表現を付与することが可能となった．多様な表現が可能となったからこそ，ユーザが所望する歌声をデザインすることを支援する研究も行われている．本研究では，歌唱表現の中でもビブラートに着目し，そのデザインを支援するインタフェースを検討する．本稿では，歌声を聴きながらビブラートデザインを行う手法を提案し，提案手法をインタフェースとして試作した．本インタフェースを構成する機能として，リアルタイムビブラートデザイン機能やデザイン対象音声のピアノロール表示機能，音声ファイルの読み込み，書き込み機能などについて説明する．最後に，提案手法のビブラートデザインの有効性について考察し，今後の展望について述べる．

1. はじめに

コンピュータを用いた楽曲制作は古くから行われており，楽曲制作に必要な機材の減少や値段の低下により，アマチュアでも楽曲制作に参入しやすくなった．それらの背景に加えて，VOCALOID [1] や UTAU [2] といった歌声合成ソフトも開発され，楽曲制作に利用されている．歌声合成ソフトは，ピアノロールという楽譜にノートと呼ばれる歌詞付きの音符を打ち込む．その後，各種パラメータを調節することで，歌声を合成する．本研究では，歌声合成ソフトによって，歌声を作り出す操作のことを歌唱デザインと定義する．

VOCALOID などの歌声合成技術・ソフトウェアの登場により，歌声合成の知識を持っていない人でも歌声を合成することが可能となった．しかし，これらのソフトウェアでユーザが所望する歌声をデザインするためには，熟練した技術を要する．その理由の 1 つとして，パラメータ調節後の歌声の予想が困難であることが考えられる．ユーザは，パラメータを調節し，歌声を合成し聴取し確認するという作業を，所望の歌声が出来るまで繰り返さなければならない．そのため，歌唱デザインを支援する研究が行われている．例として，人間の歌を真似るシステム VocaListener [3], [4] が提案されている．手本とする歌声を用意することで，所望する歌声を得ることができる．また，Sinsy [5] や CeVIO [6] などの人間らしい自然な歌声を自動的に作り出す手法も提案されている．ユーザは歌詞と譜面を入力することで，人間らしい自然な歌声を得ることができる．このように，歌声を作り出すことやそれを支援する技術開発や研究は盛んに行われており，その手法も多様である．

本研究では，歌声の自動生成ではなく，ユーザが所望の歌声を得るために実施する歌唱デザインを支援する技術開発を目指す．特に，実時間音声合成手法を用いることで，歌唱デザインにかかる作業時間を短縮する方法を提案する．従来の歌唱デザインを支援する研究について調査し，従来の研究との位置づけを明らかにする．その後，提案手法の有効性について考察する．

本稿は，以下の流れで構成される．第 2 章において，本研究でデザインの対象するパラメータと，歌唱デザイン支援について調査した関連研究について述べる．第 3 章では，関連研究に対する提案手法を示し，その機能要件について述べる．また，研究で使用した音声分析合成システム WORLD について述べる．第 4 章では，提案手法をアプリケーションとして実装した「歌唱デザインツール - Parrot」の機能と仕組みについて述べる．第 5 章では，提案手法のビブラートデザイン支援の有効性について考察をする．最後に第 6 章は結論であり，本研究において得られた成果をまとめ，今後の研究の展望について述べる．

2. 歌唱デザイン支援の関連研究

本章では，本研究でデザインの対象とするパラメータについて述べ，そのパラメータをデザインする関連研究とし

1　山梨大学　〒 400-8511, 山梨県甲府市武田 4-3-11
2　明治大学　〒 164-8525, 東京都中野区中野 4-21-1
a)　g19tk007@yamanashi.ac.jp

て，VOCALOID 4 Editor，VOCALID V Editor，VocaListener，統計的歌声合成を紹介する．その後，関連研究に対する，本研究の位置づけについて説明する．

2.1 本研究で取り扱うパラメータ

歌声合成に用いられるパラメータは複数存在する．VOCALOID を例にすると，声の高さや声の大きさ，声の息の量などが挙げられる．これらのパラメータの中で，声の高さは歌唱表現に利用される重要なパラメータである．声の高さを調整することによって，ポルタメントや，オーバーシュート，プレパレーション，ビブラートなどの歌唱表現の付与が可能となる [7]．ポルタメントとは，ある声の高さから別の声の高さへの遷移を滑らかにする歌唱表現である．オーバーシュートは，ある声の高さから別の声の高さに遷移したときに，目的の高さを通り越し瞬時的に高く，もしくは低くなる歌唱表現である．プレパレーションは，ある声の高さから別の声の高さに遷移する前に，瞬時的に目的の高さと逆に声の高さが変化する歌唱表現である．ビブラートは声の高さを上下させる歌唱表現であり，振幅と周波数のパラメータを持ち，これらは時間変動する [8], [9]．そのため，ビブラートは表現の幅があり，個人性が確認されている [10]．本研究では，ユーザが歌唱デザインにおいて個性を演出するのに利用しやすいと思われるビブラートを対象とする．

2.2 VOCALOID 4

VOCALOID 4 は，ビブラートデザインにビブラートの長さと振幅，周波数を指定する．ビブラートの長さは割合で表され，0 から 100 の数値を指定する．ビブラートの振幅と周波数は，横軸を時間とした時系列グラフに振幅と周波数を指定する．時系列のグラフであるため，振幅と周波数は時間変動させることができる．

VOCALOID 4 の利点は，ユーザがビブラートの長さと振幅，周波数の 3 つのパラメータを指定することで，詳細なビブラートデザインを行うことが可能な点である．例えば，ビブラートの振幅と周波数を時間変動させることで，ビブラートのかかり始めを穏やかにし，徐々に激しくする表現をデザインすることが可能である．本研究での目標に対する課題としては，3 つのパラメータをそれぞれ独立に指定するため，1 回のデザインに多くの時間を費やすことが挙げられる．また，デザイン後のビブラートの変化を予想するのが困難なため，所望するビブラートができるまで，デザインを何度も行う必要があることも挙げられる．

2.3 VOCALOID V

VOCALOID V は，ビブラートデザインにプリセットを用いる．ユーザは複数あるプリセットから所望するビブラートに近いものを選ぶ．その後，ビブラートの長さと振幅，周波数の調整を行う．ビブラートの長さは視覚的に表され，マウス操作で指定する．ビブラートの振幅と周波数は，2 軸のインタフェースを用いて，同時に指定する．

VOCALOID V の利点は，ユーザが所望するビブラートに近いものをプリセットの中から選んでデザインを行うため，1 回のデザインにかかる時間が短くなる点である．また，ビブラートの長さと振幅，周波数の指定が視覚的になり，デザイン後のビブラートの変化を予想しやすいことも利点である．本研究での目標に対する課題としては，プリセットが有限であるため，プリセットにない意図的に逸脱させる個性的なビブラートのデザインが困難であることが挙げられる．

2.4 VocaListener

VocaListener [3], [4] は，人間の歌を真似るシステムである．ユーザのビブラートなどの歌唱表現から歌声合成パラメータを自動推定し，推定したパラメータから歌声を合成することでユーザの歌唱を真似た歌声を得ることができる．歌声合成パラメータは，推定後手作業で調整することも可能である．歌声合成パラメータは，声の高さと声の大きさである．VocaListener2 では，歌声合成パラメータに声の音色を加えた拡張がなされている．

VocaListener の利点は，ユーザが所望する歌唱を用意することで，デザインをほぼ行わずに所望する歌声を得ることができる点である．ユーザが実際に歌って所望する歌唱を用意する場合，ユーザが歌いながらデザインを行うので，デザイン後の歌声の変化を予想しやすいことも利点である．本研究での目標に対する課題は，VocaListener の目的は，合成された歌声を目標とする歌唱に近づけることなので，目標となる歌唱をユーザ自身が用意しなければならないことである．

2.5 統計的歌声合成

統計的歌声合成は，歌詞と譜面の情報のみから自然な人間らしい歌声を自動的に合成する技術である．これにより，ユーザは歌唱パラメータのデザインをすることなく自然な人間らしい歌声を得ることができる．HMM（hidden Markov model）を用いた歌声合成技術 [11] を用いたものとして，Sinsy や CeVIO が挙げられる．また，DNN（deep neural network）を用いた音声合成技術の WaveNet [12] を歌声合成に応用したもの [13] や DNN 版の Sinsy [14] が挙げられる．これらは，自然な人間らしい歌声を得るための技術であり，ビブラートのかかった歌声も得ることができる [15]．

統計的歌声合成の利点は，ユーザが歌詞と譜面といった最小限の情報から自然な人間らしい歌声を得ることができるため，デザインを行う必要がほぼない点である．本研究での目標に対する課題としては，統計的歌声合成の目的で

ある人間らしい自然な歌声が，必ずしもユーザの所望する歌声であるとは限らないということである．

2.6 本研究の位置付け

VOCALOID 4 は，詳細なビブラートのデザインを行うことができるが，デザイン結果と合成された歌声の知覚的な対応付けが容易ではない．VOCALOID V は，プリセットを用いるのでビブラートデザインは容易であるが，プリセットに対応しないデザインは困難である．VocaListener は，目標とする歌唱を用意することで，デザインをほぼ行わずにすむが，目標とする歌唱が用意できることが前提である．また，統計的歌声合成は自然な人間らしい歌声をユーザが介在せずに得ることが目的であり，本研究の多様な表現のビブラートデザインを行うという目的とは異なる．本研究では，詳細なビブラートデザインを容易にする手法を提案することで，歌唱デザインの省力化を目指す．

3. 詳細なビブラートデザインを容易にする手法の提案

本章では，詳細なビブラートデザインを容易にする手法について述べる．まず，提案手法について述べ，提案手法のコンセプトと機能要件について述べる．その後，提案手法の実現に使用した技術について説明する．

3.1 ビブラートデザインをしながら歌声を合成，再生する手法

本研究では，ビブラートデザインをしながら歌声を合成，再生する手法を提案する．これにより，ビブラートデザインによる歌声の変化の確認が即時にできる．以下に従来のビブラートデザインの手順を示す．パラメータ調節後の歌声の変化の予想が困難なため，調節を何度も繰り返し行わなければならないという問題がある．提案手法では，パラメータ調節と歌声の合成を同時に行うことで，歌声の変化の確認を即時にでき，ビブラートデザインの省力化が可能となる．

(1) パラメータ調節
(2) 歌声合成
(3) デザインした歌声の確認

これらの条件を満足するための機能として，ビブラートデザイン後の即時の歌声のフィードバックと，ビブラートデザインを容易にするインタフェースが挙げられる．これらの機能を実現することで，ユーザは操作に対応したフィードバックを得ながら，ビブラートデザインが行える．デザインが容易になれば，作業効率も上がり，作業時のストレスを軽減することが期待できる．

3.2 音声分析合成システム WORLD

提案手法の機能要件であるビブラートデザイン後の即時の歌声のフィードバックを実現するために，高品質な音声をリアルタイムに合成する必要がある．本研究では，その機能を有している音声分析合成システム WORLD [16] を使用する．WORLD はボコーダ方式 [17] の音声分析合成システムである．WORLD による音声処理の流れを図 1 に示す．WORLD は，音声を分析し，音声から抽出した 3 つのパラメータを用いて音声を合成する．3 つのパラメータは，基本周波数（F0）とスペクトル包絡，非周期性指標である．F0 は声の高さ，スペクトル包絡は声の音色，非周期性指標は声のかすれ具合に相当する．これらの音声パラメータは独立した変換が可能である．本研究では，ビブラートデザインを行うため，声の高さに相当する F0 の変換に WORLD を利用する．

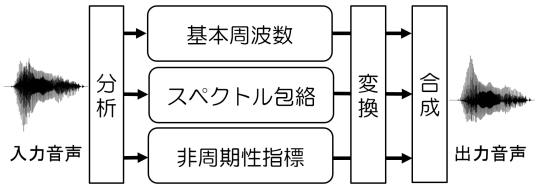

図 1 WORLD による音声処理

3.3 WORLD による実時間音声合成

WORLD は，実時間音声合成を実現するための拡張が行われている [18]．前節で説明した通り，WORLD は音声を分析し，3 つのパラメータを得る．通常，WORLD による音声合成はこれらの音声パラメータを一括で合成し，音声波形を得る．実時間音声合成は，任意のサンプル単位で音声の合成を行う．任意のサンプル単位で音声合成を行うために，実時間合成用の構造体を導入している．その構造体は，リングバッファであり，音声パラメータへのポインタを保持する．音声パラメータを逐次追加するために，音声パラメータをリングバッファにリンクする．また，合成に利用しなくなった音声パラメータへのリンクを破棄するために，現在までに合成された波形を示す時刻を記録している．

4. 歌唱デザインツール - Parrot

ビブラートデザインの省力化を行うため，ビブラートの振幅と周波数，時間の指定と同時に歌声が合成されるインタフェース「歌唱デザインツール - Parrot」の開発を行った．本章では，その詳細について説明する．

4.1 Parrot の開発環境と機能

Parrot を開発した OS は，Windows 10 Pro である．統合開発環境は，Visual Studio 2017 で，開発言語は C++である．Parrot の GUI やオーディオ出力，ファイルの読み書きなどを実装するにあたり，ROLI 社が提供するフレームワーク JUCE を使用した．Parrot は以下の機能を有する．

- 加工する歌声の音声ファイルの読み込み
- 加工する歌声の F0 軌跡をピアノロール上に表示
- 歌声を聴きながら，ビブラートデザイン加工
- デザインした歌声の再生
- デザインした歌声を音声ファイルへ書き出し

4.2 加工する歌声の音声ファイルの読み込み

今回実装した Parrot は試作のため，UTAU などで打ち込まれた歌声を読み込み，加工するという形を採用した．Parrot で歌声のビブラートデザインを行うために，デザインの対象となる歌声の音声ファイルを選択する．図 2 が Parrot の実行画面である．まず，上部の open ボタンを押下するとファイル選択のダイアログボックスが開き，加工対象となる歌声の音声ファイルを選択する．選択した音声ファイルを WORLD で分析し，歌声の加工ができる状態にする．また，Parrot が動作する音声ファイルは WAVE 形式で，モノラル音源，サンプリング周波数が 48 kHz，量子化ビット数は 16 bit である．

4.3 加工する歌声の F0 軌跡をピアノロール上に表示

歌声の音声ファイルの読み込みと分析が行われると，WORLD により，基本周波数（F0）とスペクトル包絡，非周期性指標の 3 つのパラメータが得られる．その中の F0 を利用し，その軌跡をピアノロール上に表示する．ピアノロール上に適切に歌声の F0 軌跡が表示されるために，音声ファイルの読み込みの前に図 2 の BPM，SIGNATURE に歌声のテンポと拍子を入力する．

4.4 楽曲再生中の実時間ビブラート加工

図 2 の右下にある edit ボタンを押下すると歌声が再生され，再生位置にはタイムスライダーが表示される．ユーザは歌声の再生中に，コントローラを操作し，ビブラートデザインを行う．デザインと同時に F0 軌跡のレンダリング，ビブラートのかかった歌声の合成と再生が行われる．これにより，歌声の変化が即時に知覚できる．コントローラは，マウスやペンで操作し，コントローラを押している間，歌声の再生箇所にビブラートがかかる．押す座標により，ビブラートの振幅や周波数が変化する．コントローラの縦軸の vibrato depth はビブラートの振幅を表している．座標が上に移るにつれ，振幅は大きくなる．コントローラの横軸の vibrato rate はビブラートの周波数を表している．座標が右に移るにつれ，周波数は大きくなる．

ビブラートを付与した歌声を合成するために，歌声の F0 を変換する．ビブラートの付与にあたり，コントローラからビブラート振幅 A と周波数 f_i を取得する．次に，変換対象のフレーム n のビブラートを式 (1) を用いて計算し，$F0_n$ に重畳する．T はフレーム間の時間である．最後に，変換した F0 ともとのスペクトル包絡および非周期性指標

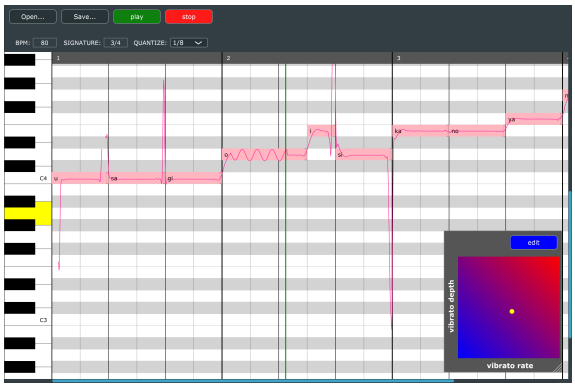

図 2 歌唱デザインツール - Parrot

を用いて音声を合成する．これにより，ビブラートが付与された歌声が合成される．

$$\Delta F0_n = A \times F0_n \times \sin\left(2\pi T \sum_{i=1}^{n} f_i\right) \tag{1}$$

4.5 デザインした歌声の再生

デザインした歌声のビブラートを確認するために，図 2 の上部の play ボタンで，歌声の再生ができる．また，stop ボタンで歌声の再生を一時停止することができる．

4.6 デザインした歌声を音声ファイルへ書き出し

デザインした音声を保存するために，音声ファイルへの書き出しを行う．図 2 の上部の save ボタンを押下するとファイル保存のダイアログボックスが開き，デザインした歌声の音声ファイルを名前を付けて保存する．デザインした歌声の音声ファイルへの書き出しは WORLD が行う．WORLD が，Parrot でデザインした歌声の F0 ともとのスペクトル包絡，非周期性指標を用いて音声を合成し，音声ファイルを出力する．

5. 考察

本章では，Parrot の機能と Parrot のビブラートデザインの有効性の検証について考察をする．

5.1 Parrot の機能

Parrot 独自の機能である歌声を聴きながらのビブラートデザイン機能を使用するために，デザイン対象とする歌声を用意しなければならない．VOCALOID などの歌声合成ソフトウェアでは歌詞と譜面情報と歌声を制御するパラメータから歌声を生成する．しかし，本研究では歌唱デザインの工程の中のビブラートデザインの支援が目的であるため，その前の段階のビブラートのかかっていない音声があるという前提で，ビブラートデザインを行う．そのため，歌声の音声ファイルを読み込み，ビブラートデザインしたものを音声ファイルに書き出す機能を実装した．これによ

り，ビブラートデザイン以外の歌唱デザインの工程を補完できると考えられる．

歌声を聴きながらのビブラートデザイン機能は，ユーザはビブラートがかかるタイミングを捉えることが容易になると考えられる．また，ユーザはデザインと同時に歌声の確認ができるので，すぐに次のデザインに移ることができ，デザイン時のストレス軽減も考えられる．さらに，ピアノロール上に表示された F0 軌跡がデザインと同時に更新されるので，耳と目の 2 つの感覚でデザインをすることができ，ユーザは従来のエディタよりも容易にビブラートデザインが行えると考えられる．

5.2 Parrot のビブラートデザインの有効性の検証

Parrot のビブラートデザインの有効性を検証するために，評価方法について考察する．Parrot の目的は，詳細なビブラートデザインを容易にすることである．それを検証するために，目標となるビブラート音声を Parrot でデザインし再現する．再現したものが目標にどれくらい近づけたかを評価するといった方法が考えられる．Parrot は，ビブラートデザイン支援インタフェースなので，インタフェースの評価としてユーザビリティテストを行うことが妥当であると考えられる．ユーザビリティ評価は有効さ，効率，満足度の 3 つの要素を評価する．有効さは，前述した通り，目標となるビブラート音声を再現し，再現したものが目標にどれくらい近づけたかを評価する．効率は，目標となるビブラート音声の再現に要した時間やデザインの回数で評価する．満足度は，目標となるビブラート音声の再現作業について，ユーザにアンケートを取ってその結果を評価する．こういったことが考えられる．また，関連研究との比較として，歌唱デザイン支援のアプローチが同じと考えられる VOCALOID 4 のビブラートデザインインタフェースに対してもユーザビリティテストを行い，Parrot の結果と比較することが，Parrot の有効性を示す方法として妥当であると考えられる．

6. 今後の展望

本稿では，詳細なビブラートデザインを容易にするために，ビブラートデザインをしながら歌声を合成，再生する手法を提案した．提案手法を「歌唱デザインツール - Parrot」として実装し，その機能について説明をした．今後，前章で考察した評価実験を行う予定である．まずは，目標とするビブラート音声は VOCALOID などの歌声合成ソフトウェアを用いて，ビブラートの振幅や周波数，長さを変えた数種類の音声を作成することを考えている．また，目標とする音声を聴いて，Parrot や VOCALOID 4 のビブラートデザインインタフェースでビブラートデザインを行う実験用のインタフェースの作成を行う予定である．

将来の展望として，ビブラート以外の歌唱表現にも対応することが挙げられる．しゃくりやフォール，ポルタメントなどの声の高さを変化させるものや，がなり声やささやき声などの音色を変化させる歌唱表現も存在する．本研究では，リアルタイムで F0 の変換を行っているが，WORLD は F0 以外の音色に相当するスペクトル包絡や声のかすれ具合に相当する非周期性指標も変換することが可能である．これらのパラメータを容易に操作できるインタフェースの開発をすることも，重要な課題といえる．

謝辞 本研究は，JST さきがけ JPMJPR18J8 の支援を受けた．

参考文献

[1] Kenmochi, H. and Ohshita, H.: VOCALOID - Commercial singing synthesizer based on sample concatenation, in Proc. INTERSPEECH 2007, pp. 4010–4011 (2007).
[2] 飴屋/菖蒲 : 歌声合成ツール UTAU サポートページ, 入手先 <http://utau2008.web.fc2.com/> (参照 2019-05-27).
[3] Nakano, T. and Goto, M.: VocaListener: A singing–to–singing synthesis system based on iterative parameter estimation, in Proc. SMC 2009, pp. 343–348 (2009).
[4] Nakano, T. and Goto, M.: 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. ICASSP 2011, pp. 453–456 (2011).
[5] Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y. and Tokuda, K.: Recent development of the HMM–based singing voice synthesis system - Sinsy, in Proc. Speech Synthesis Workshop, pp. 211–216 (2010).
[6] (c)CeVIO.: CeVIO Official Web, 入手先 <http://cevio.jp/> (参照 2019-05-27).
[7] Saitou, T., Unoki, M. and Akagi, M.: Development of an F0 control model based on F0 dynamic characteristics for singing–voice synthesis, Speech Communication, Vol. 46, pp. 405–417 (2005).
[8] Bretos, J. and Sundberg, J.: Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos, TMH-QPSR, KTH, Vol. 43, No. 1, pp. 37–44 (2002).
[9] Prame, E.: Measurements of the vibrato rate of ten singers, STL-QPSR, Vol. 33, No. 4, pp. 73–86 (1992).
[10] Migita, N., Morise, M., and Nishiura, T.: A study of vibrato features to control singing voices, in Proc. ICA2010, PaperID:164, Sydney, Australia, Aug. pp. 23–27 (2010).
[11] 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. 405–417 (2005).
[12] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W. and Kavukcuoglu, K.: WaveNet: A generative model for raw audio, CoRR, arXiv preprint arXiv1609.03499 (2016).
[13] Blaauw, M. and Bonada, J.: A neural parametric singing synthesizer, arXiv preprint arXiv1704.03809 (2017).
[14] Hono, Y., Murata, S., Nakamura, K., Hashimoto, K., Oura, K., Nankaku, Y., Tokuda, K.: Recent development of the DNN–based singing voice synthesis system - Sinsy, in Proc. APSIPA ASC 2018, pp. 1003–1009 (2018).
[15] 山田知彦, 武藤聡, 南角吉彦, 酒井慎司, 徳田恵一 : HMM に基づく歌声合成のためのビブラートモデル化, 情報処

理学会研究報告, Vol. 2009-MUS-80, No. 5, pp. 309–312 (2009).
[16] Morise, M., Yokomori, F. and Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, IEICE transactions on information and systems, Vol. E99-D, No. 7, pp. 1877–1884 (2016).
[17] Dudley, H.: Remaking Speech, J. Acoust. Soc. Am., Vol. 11, No. 2, pp. 169–177 (1939).
[18] 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, Vol. 2016-MUS-112, No. 20, pp. 1–6 (2016).