公開

研 究 終 了 報 告 書

「Human-in-the-loop 型歌唱デザインの開発」

研究期間: 2018 年 10 月～2022 年 3 月
研 究 者: 森勢 将雅

1. 研究のねらい

本研究では，高さなどの歌声を構成するパラメータを作りこむ歌唱デザインを計算機がサポートする，Human-in-the-loop 型歌唱デザイン技術の研究開発を目指す．これは，歌詞と譜面に基づき人間と等価な品質の歌声を自動生成する，現在主流である歌声合成研究の次のシーズ発掘を見据えた研究である．現在の歌声合成は VOCALOID や CeVIO などがすでにソフトウェアとして実用化されており，ユーザは歌詞と譜面により自然な歌声を生成可能である．一方，人間と等価な品質であることは，必ずしも歌唱デザインが不要であることを意味せず，生成後に自身のイメージに沿うようにエラボレーションを作りこむ必要がある．この作業を計算機により支援し省力化する基礎技術の開発が，本研究の狙いである．

一連の研究は，主に 3 つの研究領域に分割して実施する．第一のテーマは，歌唱デザインを支援するインタフェースの実現である．歌唱デザインにおける課題の 1 つに，デザイン結果を確認するには合成・再生にかかる時間，すなわちデザイン終了から結果を確認するまでのタイムラグが挙げられる．デザイン作業は 1 つの音符に対し複数回行う可能性を考慮すると，タイムラグは作業時間の増加や作業におけるストレスに繋がる．本研究では，この問題に対し「聴きながらデザインする」インタフェースで解決を図る．

第二のテーマは，ユーザがデザインしたパラメータに基づき自然な歌声を生成する歌声合成技術の実現である．歌声のパラメータを加工した場合，一般には加工の程度に依存して音質が劣化する．本研究では，従来用いられる歌詞と譜面から自然な歌声を生成する Deep neural network（DNN）に加え，デザイン結果に基づいて自然な歌声を生成する DNN を構築し，高い品質を維持したまま歌声を加工する手段を提供する．

最後のテーマは，人間が知覚する歌声の自然さを計測するモデルの構築である．本研究では，歌唱技巧の 1 つであるビブラートと発声タイミングのズレに絞って検討を進める．それぞれが自然と判断される範囲を明らかにしインタフェースに組み込むことで，ユーザの作業効率化を図る．以上の研究により得られた成果物を社会へ還元し，コンテンツ制作のフリーソフトや製品開発に繋げ，歌声合成に関する研究・開発者，コンテンツ制作者人口を増やすことも，本研究の狙いに含まれる．

2. 研究成果

（1）概要

本研究では，歌唱デザインを省力化するための一連の検討を実施し，(1) 実時間で歌声をデザインする機能を組み込んだ歌唱デザインインタフェースの研究，(2) デザイン結果から自然な歌声を生成する Deep learning の研究，および(3) 歌唱技巧の自然性を判断するモデル構築の研究に取り組んだ．ここでは，それぞれの研究領域に関する成果概要を報告する．当初計画していない成果や波及効果については詳細の節で述べる．

歌声の実時間デザインについては，研究者が開発した音声分析合成システムをベースに発展させる形で実現した．本研究では，様々な歌唱技巧から多くの歌声合成ソフトウェアで制御対象とされるビブラートに特化して，歌声を聴きながらビブラートを制御できるインタフェースを構築した．このインタフェースを主観評価し，既存のエディタのみ用いた場合よりも作業時間が短縮され，ユーザビリティも改善できることを示した[論文 3]．加えて，精密な操作を実施できる既存のエディタが有効な場面があることも明らかになったため，実時間デザインと既存のエディタを組み合わせることで，目的とする歌声をより精密にデザインできることを示した．

上述のインタフェースで制御できるパラメータはビブラートのピッチ軌跡のみであり，音色や大きさのパラメータが変化しない．そのため，合成結果の品質を向上させるためには高さに連動してパラメータを変化させる必要がある．本研究では，デザインされたビブラートのピッチ軌跡を入力として，音色と大きさを生成する DNN を構築することで，この問題の解決を目指した．検討の結果，提案するインタフェースで制御したピッチ軌跡に限定することで，自然な音色と大きさを生成できることが確認できた．一方，ユーザがフリーハンドで制御したピッチ軌跡では，制御の度合いにより大幅な劣化が生じることも確認された．

歌声の知覚については，上述の研究で扱うビブラートに加え，背景楽曲と発声タイミングのずれがどの程度まで許容されるかについての主観評価も実施した．ビブラートの自然性については，複数の二次元ガウス関数を 3 個重ね合わせたモデルにより，主観評価と比較して 5% 未満の誤差で概ね近似できることが示された[論文 2]．タイミングについては，正確な時刻に対し±30 ms 以内のずれであれば大きな違和感とはならないことが明らかとなった．

（2）詳細

研究テーマ1「歌唱デザインインタフェース」

本テーマでは，特に歌声をデザインした結果がリアルタイムで確認できず，合成，再生というタイムラグが発生することの解決を目指した．歌唱技巧は複数あるが，その中でも音の高さを小刻みに振動させるメジャーな技巧であるビブラートに着目し，再生中にリアルタイムでビブラートを制御できるインタフェース「Parrot」を提案した[論文 3]．Parrot では，図1の右下に示されるエディタにより，ビブラートを加工する．具体的に，ビブラートは正弦波的に高さを振動させるため，正弦波の周波数（速さ）と振幅（深さ）をエディタの横軸と縦軸として与え，ユーザがクリックした座標で定まる速さ・深さのビブラートが，再生中の歌声に付与される．本エディタを

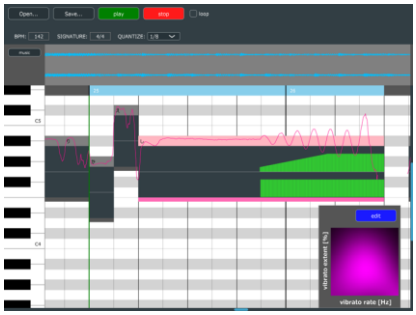

図 1：Parrot のスナップショット．

用いることで，ユーザがビブラートをかけるタイミングを再生中に決めてリアルタイムでパラメータを変えながら制御することが可能となる．また，エディタには紫色でグラデーションがかかっているが，これは後述する歌声知覚の研究で得られたビブラートパラメータと自然性との関係を示している．明るい紫ほど自然と判断されるため，ユーザはこの色を手掛かりに自然なパラメータの範囲から所望の特性を定めることが可能となる．

本エディタについて，デザイン後に再生して結果を確認する既存のエディタを用いた場合との比較実験を実施した．実験では，お手本となるビブラートを模擬するタスクを被験者に課し，作業にかかる時間やユーザビリティ，作りこんだビブラートパラメータがどの程度お手本に近づいているかなど，主観・客観的な指標により有効性を検証した．実験の結果，提案するエディタは，作業時間の短縮とユーザビリティが有意に向上できることが確認できた．作りこんだパラメータの精度については，既存のエディタと有意な差が認められなかったが，これはどちらのエディタでも同程度に精密なデザインが可能であったことを意味する．

本テーマでは，ビブラートに加えて，アイドルが利用する語尾を跳ね上げる歌唱技巧も制御対象としてインタフェース開発に取り組んだ．これは，第2年次に構築した歌唱データベースに歌声を提供した声優がアイドル歌手であり，跳ね上げ表現を多量に含んだデータベースが構築できたため，新たに取り組むことにした研究である．図2は，実装したインタフェースのスナップショットであり，赤の実線がピッチ軌跡である．図からも，語尾でピッチが跳ね上がっていることが確認できる．語尾跳ね上げは，前述の歌唱データベースから跳ね上げを含む区間を抽出し，シグモイド関数により近似できることに着目して実装した．具体的には，語尾を跳ね上げる高さと急峻さをシグモイド関数のパラメータで表現し，ユーザは開始地点，跳ね上げる程度と急峻さの目安をマウス操作で与えることで，シグモイド関数に基づき跳ね上げを付与することが可能となる．

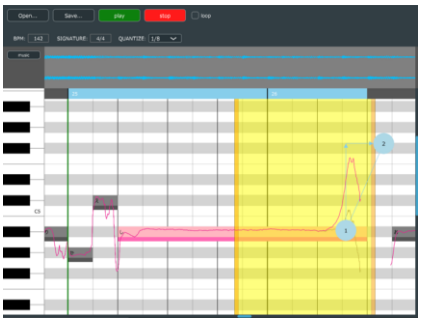

図2：実装した語尾の跳ね上げ機能のスナップショット．

本インタフェースについてビブラートデザインと同様の手順で評価したところ，合成結果の音質についてはフリーハンドでデザインした場合と比較して有意に向上することが確認できた．一方，作業時間やユーザビリティについての全体的な評価では有意な差が認められなかった．これは，お手本となる歌声の跳ね上げ幅が小さい場合のデザインが困難であったことに起因する．跳ね上げ幅が大きい場合に限定して解析すると，作業時間の短縮に寄与することも確認できた．このように，歌唱デザインインタフェースに関しては，ビブラートと語尾の跳ね上げに対する作りこみを支援する機能を提案し，有効性を示せたといえる．

研究テーマ2「デザイン結果から自然な歌声を生成する Deep Neural Network」

テーマ2では，テーマ1のインタフェースで制御したビブラートのピッチ軌跡から，より自然な歌声を生成する Deep Neural Network (DNN)を設計するための検討を実施した．テーマ1ではビブラートに対応するピッチ軌跡をリアルタイムで与えるのみで，音色や大きさについては制御していない．したがって，リアルタイムで結果の目安を知るためには十分であるが，最終的に自然な歌声を得るためには，その結果に基づいて音色と大きさを制御する必要がある．本

テーマでは，この課題を解くための DNN 構築を目指した．本来であれば，ユーザがビブラート以外にも様々な歌唱技巧を表現する可能性を考慮して，任意のピッチ軌跡から自然な歌声を生成できることが望ましい．しかしながら，予備検討の結果，人間が発声できないような極端に変化するピッチ軌跡では音声そのものが破綻することが確認されたため，本テーマではテーマ1でデザインしたビブラート歌唱をより洗練させるDNNに問題を限定して実施することとした．

構築した DNN では，言語特徴量に加えてユーザがデザインしたピッチ軌跡を入力として与え，出力をピッチ軌跡，音色（スペクトル包絡），声の掠れの程度（非周期性指標）として学習した．これら 3 種類はチャネルボコーダに基づく音声パラメータであり，3 種の音声パラメータから人間とほぼ等価な音声が合成できることは，他の研究ですでに示されている．出力にピッチ軌跡を含めることは，ユーザがデザインしたピッチ軌跡は正弦波的な振動として与えているため，そこから人間らしいビブラートのピッチ軌跡に変換するために必要である．

DNN により処理した波形の一例を図3に示す．図の上段が元の歌声波形にテーマ 1 のインタフェースでビブラートを付与したものであり，下段が提案する DNN により音声パラメータを生成した結果である．提案する DNN を用いることで，エンベロープに変化が生じ，語尾にかけてパワーが滑らかに減衰する特性が付与されていることを確認できる．こうして得られたビブラート歌唱の品質を主観評価で確認したところ，平均的に品質は改善することが確認できた．ただし，いくつかのビブラート歌唱では，品質が変化しないことも明らかとなった．この歌声は，学習に用いた歌手が発声できるビブラートから大きく逸脱した場合であり，自然と知覚されるビブラートの範囲であれば，概ね良好に学習可能と解釈している．本テーマの成果は，条件やパラメータの範囲に制約があるため，ユーザがデザインした結果から自然な歌声を生成するという目的を達成できたが，効果は限定的と考えている．様々なピッチ軌跡においても自然な歌声を生成できる技術の確立は，今後の重要な課題である．

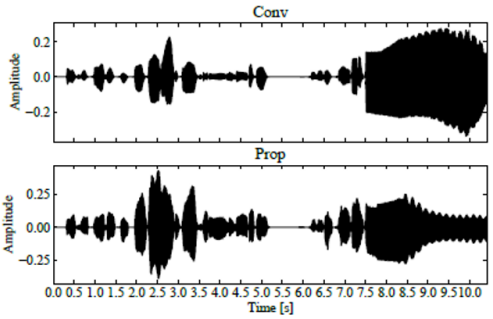

図 3 ：ビブラートのピッチ軌跡に基づく波形の制御結果

研究テーマ3「歌声の自然性を計測するモデルの構築」

歌唱デザインではビブラートを扱っていることから，本テーマでは，ビブラートを構成する速さ・深さ（ビブラートパラメータ）がどの程度であれば自然と知覚されるかについて範囲を明らかにすることを狙う．加えて，発声のタイミングのズレは大きな違和感として知覚されることから，背景楽曲と発声タイミングのズレが違和感として知覚されない範囲についても検討した．

ビブラートの知覚実験では，様々なビブラートパラメータを有する歌声を評価に用いる必要があるため，VOCALOID を利用して速さ・深さを系統的に変化させた歌声を用意した．複数の音高，複数の歌手でも歌声を生成することで，音高や歌手に対する依存性があるかについても同時に検討している．最終的に音高依存性が観測された場合を想定し，評価結果を数理モデルとしてパラメータ表現し，音高の差による自然性の範囲をパラメータ補間で自然に実施す

ることを目指した.

実験結果の一例を図4に示す. 図の横軸はビブラートの深さ, 縦軸は速さに対応し, 主観評価結果を目視で観察した結果釣鐘状であることから, 2次元ガウス関数の重ね合わせにより近似した. 図4は3個の混合で表現し, 主観評価との差は4%程度であった[論文 2]. 図の等高線上に記載された数値は予測される自然性の数値であり, 4が最も自然に対応する. 本結果の数値に色を割り当ててエディタの背景色としたものが, 図1のビブラート制御エディタである. このように, 歌声知覚の成果をインタフェースに組み込むことができた.

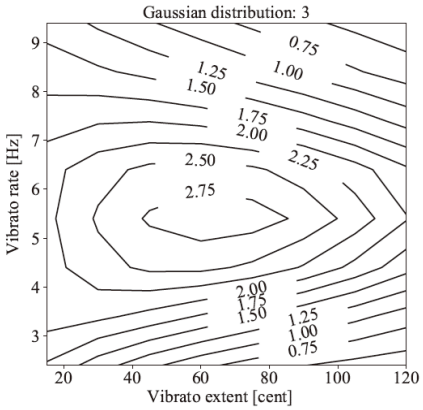

図 4:ビブラートパラメータと自然性を表すマップ.

背景楽曲と発声タイミングのズレについても同様の流れで主観評価を実施した. こちらは歌声の高さや音色を加工する必要がないため, 歌唱データベースに収録されているプロの歌声をタイミングが合っている真値とみなし, そこから時間をずらして背景楽曲と重畳して再生することで, 知覚する違和感を評価した. 評価の結果, こちらは早い場合と遅い場合とで違和感と知覚する範囲に概ね差はなく, ±30 ms 以内であれば大きな違和感として知覚されないことが確認できた(図 5). 本結果については, 現状の DNN で生成した歌声の発声タイミングが概ね 30 ms 以内であったため, タイミング修正のインタフェース開発などは実施していない. また, 図5はテンポがバラバラの 10 曲に対して得られた結果を平均したものである. 曲のテンポや各音符の平均的な長さについての相関を解析したところ, 楽曲のテンポに対する相関は弱く, 音符の平均的な時間長と相関することが確認された.

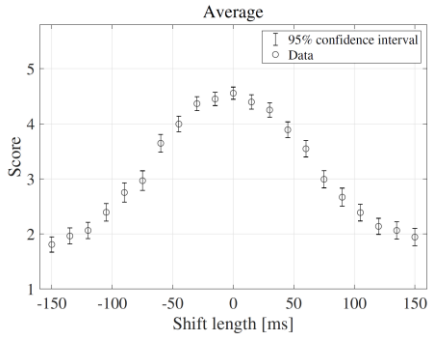

図 5:タイミングのズレと自然性のスコアとの関係.

その他の成果

本研究では, 歌声合成や歌声知覚を研究するための資料として, プロの声優兼アイドル歌手による 50 曲(有声区間が約 1 時間)を収録した歌唱データベースを構築した[論文 1]. 研究の波及効果を高めるため, 本データベースは事前に承諾を得て, 研究用途であればフリーで利用できるようにし, 商用利用についても契約を結べるようライセンスを整備した. 本データベースは NEUTRINO と呼ばれるフリーソフトの公開や, 音声創作ソフトウェア CeVIO に商用利用として組み込まれるなど, すでに利用された実績がある.

前述の歌唱データベースは著作権の問題で利用に制限がかかるため, 同程度の規模で全てオリジナルの楽曲で構成される歌唱データベースも構築した. これらの活動を通じて歌声や音声データベースを構築しライセンスを整備する手順を確立した. 現在採択されている科研費基盤 A の音声デザインのプロジェクトにおいても, 引き続き協力体制を敷いている.

3. 今後の展開

本研究の成果の一部（歌唱データベース）は，すでに国内外で多数の研究者に利用されている．具体的には，NEUTRINO と呼ばれるフリーソフトの開発・公開や，音声創作ソフトウェアである CeVIO に採用され歌声合成ソフトウェアとして製品化されたことは，社会実装に関する波及効果として挙げられる．公開した歌唱データベースの波及成果には，本研究で制作したことをきっかけにクラウドファンディングが企画されたことも挙げられる．本企画は十分な資金を集めて成功した（研究代表者も制作チームに参画している）など，歌声合成に関する活動は非アカデミア領域にも波及している．

歌唱デザインの基盤技術については，歌声を再生中に実時間で制御し，それを DNN で整えるという人間・計算機の協調という観点を提案した．高品質な音声合成システムを用いたリアルタイム制御のアイディアは，現在他のインタフェース開発でも利用されつつある．このように，本研究は，歌唱デザインという領域の開拓に一定の貢献があるといえる．DNN によりデザイン結果の品質を向上させる技術と，提案するような歌唱デザインインタフェースを組み合わせた製品も，技術的には数年以内に発売できる水準に達すると考えている．

歌声や音声の知覚については，現在いくつかの大学や研究機関と共同研究契約を結び，心理学分野と連携した研究を推進している．こちらは基礎研究のフェーズであり，社会実装の話へ発展するには，数年以上の時間はかかると思われる．ただし，現在の音声解析が，音声認識のようなテキスト情報を取り出す研究や，個人識別，感情認識が中心である一方，歌声のうまさや魅力など，より個人差が大きく複雑な問題を扱う研究領域の開拓につながることが期待される．心理学会でのワークショップや研究技術を紹介するセッションに参加しており，心理学分野において音声・歌声知覚に関する研究事例は，少しずつだが増えている．

4. 自己評価

本研究の目的には，主に歌声の実時間デザイン技術とそれを実現するインタフェース，ユーザがデザインした結果に基づく自然な歌声の生成，歌声の知覚モデルの構築が挙げられる．以下では，それぞれについての達成度合いについて説明し，最後に研究遂行や成果の波及効果についての評価を述べる．

研究目的については，研究成果で述べたとおり，大筋期待される成果を達成できたと考えている．実時間デザインについては，ビブラートの制御だけではなく，アイドル歌唱の表現を付与するなど，当初の計画を超えた成果が得られた．一方，ユーザがデザインした結果に基づく歌声生成では，ビブラートデザイン結果のみ有効に機能するが，ユーザがフリーハンドでデザインした逸脱した軌跡からは機能しないなど，限定的な成果となった．これは，自由なデザインだとDNNは大きく破綻した結果を出力するという，本研究の結果がベースにある．歌声のタイミング制御の研究では，自然と判断される範囲を明らかにしたが，この機能はインタフェースに組み込んでいない．これは，現状の DNN 歌声合成は，タイミングについてはほぼ自然と判断される範囲で出力されており，実質的に調整する機会がなかったことに由来する．自然と判断するビブラートパラメータの範囲については，インタフェースに組み込むことでユーザの作業を手助けするエディタとして統合することができた．このように，当初の計画以上の成果が得られたテーマ，期待された結果が得られなかったテーマなどが混在しており，全体としては各研究テーマである程度の成果が得られたと考えている．

研究の進め方については，第2，3年次は研究補助者3名，学生2名の体制で進めており，複数のテーマを並行し，効率的に進めることができた．予算の執行に関しては，特に3年次からはコロナ禍により学会がオンライン参加となったため，旅費に関する計画変更を余儀なくされた．旅費に関しては，研究に利用する歌唱データベースの品質向上に費やすなど，柔軟に執行計画を変更することで対応できたと自己評価している．

研究成果の波及効果については，今後の展開で説明したとおり，第2年次に公開した歌唱データベースは幅広く利用されている．歌唱データベースはライセンスを調整し研究用途で使いやすく，かつ商用利用もできるようにしたことで，多くの研究・開発者に利用されたと判断している．また，研究成果を論文だけではなく，音声提供者側と契約し，利用に関するライセンスまで整備したことは，直接的な成果ではないものの，今後同様のデータベースを構築する際の基盤を作ったと自負している．音声や歌声の知覚については，さきがけ研究がきっかけで心理学分野との共同研究がスタートしている．また，さきがけ研究は，歌声に限定せず音声全般のデザイン技術の構築を目指す科研費基盤 A の着想にも繋がっている．このように，さきがけ研究を基盤とした研究プロジェクトがアカデミア・非アカデミアで進んでいることも，重要な波及効果であると考えている．

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：9件（査読付き国際会議を含めると合計14件）

1. I. Ogawa and M. Morise: Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs, Acoustical Science and Technology, 2021, vol. 42, no. 3, pp. 140–145.
本論文は，機械学習に基づく歌声合成において必要不可欠な，歌唱データベースの構築についてまとめたものである．歌声は著作権の制約があり一般への公開が困難であったが，2019 年に著作権第 30 条が改定され，機械学習向けに制約が緩められた．この条件緩和を受け，ジャンルを J-POP に絞り歌唱データベースを構築した．加えて，音素バランスや譜面における高さやテンポ，各音符の長さなどの統計的性質を明らかにした．
2. T. Miyazaki and M. Morise: Building a measurement model for simulating naturalness of vibrato based on subjective evaluation, IEICE transactions on information and systems, 2021, vol. E104–D, no. 4, pp. 521–525.
本論文では，高さを振動させた歌唱技巧であるビブラートに着目し，振動の速さと深さがどの範囲であれば，人間にとって自然と知覚されるかを主観評価に基づき数理モデルを構築した．様々な速さ・深さのビブラートを生成するため既存の歌声合成ソフトウェアを用い，主観評価によりモデル化に必要なデータを得た．その後，データ結果から速さ・深さに対する 3 つの二次元ガウス関数による混合モデルにより，主観評価との誤差が 4%程度で自然性を計測できることを明らかにした．

3. 小野雄大, 森勢将雅：Parrot: 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, 第 27 回インタラクティブシステムとソフトウェアに関するワークショップ, 2019, 6-page.
本論文では, ビブラートを再生中に制御することを実現するインタフェース「Parrot」を提案し, 既存のビブラート制御エディタと比較して有効性を明らかにした. 一般的な音声や歌声のデザインでは, エディタにより編集し, その後合成・再生することで結果が明らかになる. Parrot ではこのタイムラグを削減し, ビブラートデザインの省力化を目指す. インタフェースを実装し, 作業時間やユーザビリティに関する評価から, 既存のエディタよりも作業時間を短縮し, かつ高いユーザビリティが達成できることを示した.

（2）特許出願

該当なし

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

1. 森勢将雅：さて, そろそろ本気を出して歌唱デザインを研究してみようか, CEDEC2021, Aug. 24, 2021.（招待講演）
2. 東北きりたん歌唱データベース（著作物）：https://zunko.jp/kiridev/login.php
3. No.7 歌唱データベース（著作物）：https://voiceseven.com/7dev/login.php
4. 森勢将雅：嗜好に着目した音声・歌声研究のエンタテインメント応用, ViEW2020 ビジョン技術の実利用化ワークショップ, Dec. 3, 2020.（基調講演）
5. 森勢将雅, 藤本健, 小岩井ことり, レアなモーラを含む日本語歌唱データベースの構築と基礎評価, FIT2021（第 20 回情報科学技術フォーラム), pp. 59-64, Aug. 25, 2021.（学会発表）
6. The 2021 EURASIP-ISCA Best Paper Award (Speech Communication Journal).（受賞）
7. FIT2021 第 12 回情報科学技術フォーラム, FIT 船井ベストペーパー賞.（受賞）