様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　１８　日現在

機関番号：　１３５０１
研究種目：　挑戦的萌芽研究
研究期間：　2016～2017
課題番号：　１６Ｋ１２５１１

研究課題名（和文）「自然な非人間性」に着目した新たな歌唱デザイン論の研究

研究課題名（英文）Study on new vocal design focusing on naturally dehumanized singing

研究代表者
森勢　将雅（MORISE, Masanori）

山梨大学・大学院総合研究部・助教

研究者番号：６０５１００１３

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：VOCALOIDを代表とする歌声合成ソフトウェアが広く一般に普及するにつれ，計算機による「人間的」な歌唱を目指す数多くの取り組みがなされてきた．一方，Auto-Tuneなどのソフトウェアを用いた「非人間的」な歌唱もコンテンツとして利用されている．ここでは，コンテンツとしての自然さと非人間性を両立する歌声が存在するか確認するため，人間性を制御する加工法について研究に取り組んだ．実験の結果，提案法により，人間の歌声が有する揺らぎ成分を除去するという従来のアプローチだけではなく，誇張させた場合でも一定の自然さを保ちつつ非人間的な歌声を生成できることを確認した．

研究成果の概要（英文）：Vocal design algorithms for approximating the human singing have been proposed with the growth of commercial software such as VOCALOID. On the other hand, there are music contents by the dehumanized singing, and for this purpose, applications such as Auto-Tune are generally used to remove the human-like feature in singing.
This study proposes the vocal design algorithm to output the dehumanized and natural singing. In the proposed algorithm, we first proposed the speech analysis/synthesis algorithm. And then, we propose an algorithm for exaggerating several features such as fluctuation of fundamental frequency. We carried out subjective evaluations to verify the effectiveness of the proposed algorithm. The result suggests that the exaggeration can synthesize the singing with a certain level of naturalness and humanness.

研究分野：　音声情報処理

キーワード：　感性情報学　歌声情報処理　統計的歌声合成

１．研究開始当初の背景

歌声合成に関する研究は，VOCALOID を代表とする歌唱合成ソフトウェアが普及してきた今日において，国内外の学会で歌声を対象としたセッションが組まれるなど加速度的に事例が増えている．研究事例の増加に伴い人間に近い歌唱合成技術が発展し，ビブラートなどの歌唱技巧を合成することも可能になりつつある．その一方で，Perfume など一部のアーティストは，ソフトウェアを用いて人間の歌声らしい表現（本研究ではこれを歌声の人間性と呼称する）を損なわせる加工により，独自の表現を追求している．

例えば，漫画では，ジャンルに沿った様々なデフォルメ表現が生まれており，適切にデフォルメされた絵は，実写とは異なる感動を与える．その一方，歌声のデフォルメ表現を追求する研究の事例は少なく，単にクオリティの低い合成歌唱とデフォルメされた合成歌唱の知覚的な違いも明らかにはされていない．歌詞と譜面から人間に近い歌声を生成する統計的歌声合成技術では，人間の歌声とほぼ等価な結果を生成することが可能である．統計的歌声合成では，事前に多数の歌声を用いて学習することを要求するが，学習データが不十分な場合，人間の歌声とは異なる音色となる．これは，デフォルメされた歌声とは異なり，単に品質の低い歌声であると評価される．画像でも同様に，実写から遠いことだけがデフォルメの条件ではなく，その中でも魅力的なもの，魅力的ではないものが混在する状況である．

２．研究の目的

本研究は，歌声合成において，計算機でのみ実現可能な歌唱表現の追求を目標と設定した．これは，単に歌声の表情を劣化させるだけではなく，「人間らしくはないが自然な歌声」が知覚的に存在するのかを明らかにする検討が含まれる．現在の歌声合成は，人間の歌声を模倣するための研究であり，例えば統計的歌声合成においても，多数の人間の歌声からもっともらしい歌声を生成する．この場合，もっともらしい歌声とは学習データの平均的な歌声，すなわち人間の歌声であるが，歌声のデフォルメを研究の目的とする場合，デフォルメをどのように評価するべきかという新たな問題が浮上する．

本研究では，歌声のデフォルメを行うため，歌声の人間性に注目した知覚モデルの構築を目指す．人間の声を構成するパラメータは膨大な数となるため，ここでは，その初歩的な検討として，時間的な揺らぎに着目した研究に特化する．人間の歌声の揺らぎ特性を系統的に変化させる信号処理技術を提案し，揺らぎの程度と知覚する人間性について心理実験を行うことで検討を進めた．

３．研究の方法

本研究には，音声の品質を劣化させること無く加工するためのツールが必要であり，デフォルメ歌唱の生成を考えた際には，統計的歌声合成に関する検討も必要となる．本研究では，研究代表者が開発した音声分析合成システム WORLD を基盤とし，WORLD そのものの品質を向上させることに加え，(1) 品質を損なわずに加工可能なように改良する研究，(2) 音声の人間性に関する特徴を変化させる信号処理技術の開発，(3) 提案する信号処理技術を活用した心理実験，(4) 統計的歌声合成への応用についてそれぞれ検討を進めることとした．

４．研究成果

本研究により，音声分析合成基盤の拡張と，音声特徴量のうち，人間性知覚に関するものとして，基本周波数（高さ）とスペクトル包絡（音色）の時間的な揺らぎが影響することが明らかとなった．さらに，各パラメータの揺らぎを誇張した場合の実験も行った．統計的歌声合成については，スペクトルと基本周波数のモデリングによる歌声合成と，統計的歌声合成に必要なデータベース構築について検討した．それぞれについて，成果の概要を示す．

（１）音声分析合成 WORLD の拡張と，人間性を段階的に変化させる方法の提案

本課題の成果として，音声分析合成システム WORLD を改良した．この改良により，従来同様の実験を行う際のデファクトスタンダードであった STRAIGHT を上回る性能を達

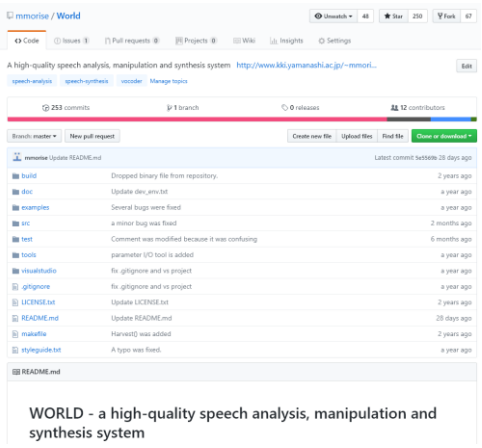

図 1：GitHub に公開した WORLD の Web ページ．

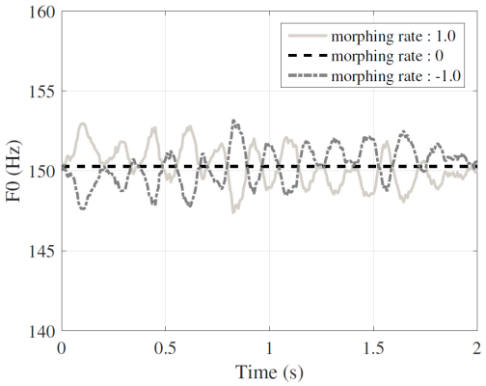

図 2：音声モーフィングによる基本周波数中の揺らぎの系統的変化.

成した[論文 1, 3]. WORLD は GitHub 上に公開し，論文の被引用数も 2018 年 6 月現在で Google Scholar Citations で 80，GitHub では 60 以上の派生プログラムが開発されている（図 1）[その他 1]. 演算を効率的に実施するため実時間で合成するための実装法も実現した[学会発表 5].

次いで，WORLD を活用し，基本周波数，スペクトル包絡に含まれる時間的な揺らぎを系統的に変化させるための信号処理技術を開発した．これは，初歩的な検討として，話声のように基本周波数が短時間で大きく変化し，また母音と子音が混ざり合うような音声ではなく，持続発声した単独音の歌声を対象とした方法である．本手法では，まず元音声の基本周波数を推定し，平均値を算出する．全有声区間の基本周波数をこの平均値に置き換えることで，時間的な揺らぎが全く含まれない音声が生成可能となる．この音声の基本周波数と，入力音声の基本周波数について重みを付けて平均を算出することで，揺らぎ特性を段階的に制御することが可能となる．スペクトル包絡については，各周波数ビンについて同様に処理することで，揺らぎを制御できる.

本手法を用いることで，基本周波数，スペクトル包絡の時間的揺らぎを独立して与えることができる．また，元音声側の重みを負値にすることで反転した揺らぎを与えることができ（図 1），1 以上にすることで揺らぎを誇張することが可能となる．本研究では，この方法を用い，基本周波数とスペクトル包絡のどちらが人間性の知覚に支配的であるか，誇張した場合は，どの程度まで人間性が保たれるかについて検討した.

（2）歌声の人間性に着目した心理実験

本手法により加工した音声を用いて主観評価実験を実施した．実験には分析合成音をリファレンスとした MUSHRA 法を用い，詳細な差を検出できるように防音室でヘッドホンを用いた評価法とした.

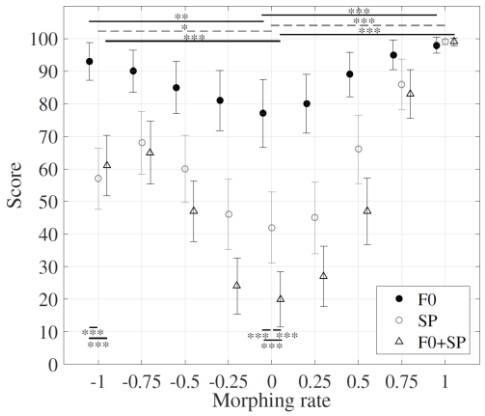

図 3：基本周波数とスペクトル包絡の揺らぎが人間性知覚に与える影響．横軸が負の場合，反転した変化を与えている.

第一の実験[学会発表 3]では，基本周波数の揺らぎの係数を-1 から 1 までとすることで，基本周波数とスペクトル包絡の揺らぎにミスマッチを与え，それが知覚する人間性にどのような影響を与えるか検討した．この結果は，図 3 に示すように，揺らぎが全く含まれない横軸が 0 の場合の人間性が最も低く，負の重みであってもある程度人間性が保たれることが示された．基本周波数とスペクトル包絡では，スペクトル包絡の影響が支配的である．また，基本周波数とスペクトル包絡を同時に制御することで，さらに人間性を低下できることも示された．一般的に，ソース・フィルタモデルでは基本周波数とスペクトル包絡の相互作用の存在が示されているが，基本周波数の時間的な揺らぎ特性に関しては，ある程度の揺らぎがあれば相互作用を気にしなくとも人間性が保たれるという結果が得られた．ただし，重みが-1 と 1 とのスコアには差が生じており，特にスペクトル包絡を変化させた場合の差が顕著であることから，揺らぎ特性の相互作用そのものは存在するといえる.

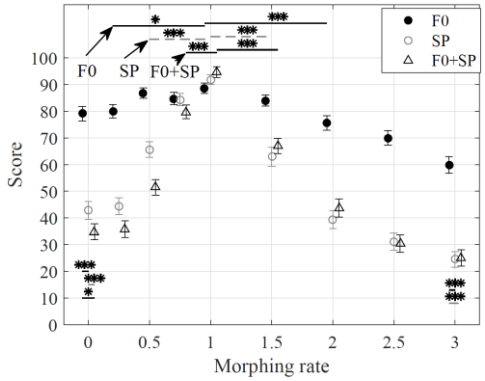

図 4：基本周波数とスペクトル包絡の揺らぎが人間性の知覚に与える影響．横軸が 1 以上の場合，変化を強調させている.

第二の実験として，1 以上の重みを用いることで揺らぎを誇張した歌声から知覚する人間性の評価を実施した[学会発表 1]．この実験は，第一の実験と同様の条件で実施した．図 4 からも明らかに，誇張させた場合の人間性についても，負の重みを与えた場合と同様に人間性が低下する傾向が得られた．こちらの結果も，基本周波数とスペクトル包絡とでは，スペクトル包絡の影響が支配的であることが確認された．一方，重みが 0 の際に観測されたスペクトル包絡と基本周波数を組み合わせた人間性が最も低いという傾向は得られず，スペクトル包絡のみの揺らぎを誇張した場合と同程度の違いとなった．

（3）統計的歌声合成への応用

統計的歌声合成への応用に向けて，まず必要となる，歌声合成のための歌唱データベースの構築を実施した[学会発表 2]．統計的音声合成を実現するための音声データベースでは，音素の出現確率を揃える，所謂音素バランス文が利用される．歌声合成については，高さそのものが音楽性に関連するため，音素バランスだけではなく，楽譜情報のバランスも考慮する必要がある．そこで，統計的歌声合成のための歌唱データベースの構築に向けて，楽曲をフレーズ単位に分割し，エントロピーに基づいて音素・音高・音高差・音長・音符の小節内の位置・テンポの 6 つのコンテキストに基づいてバランスを取ったフレーズセットを構築した．

従来の歌声データベース（CONV_DB）は，51 曲計 20978 モーラのセットである．構築したデータベースは，フレーズ単位のため曲数はないが，同程度の規模となるよう 20966 モーラを選出した．構築したデータベースを既存のデータベースと比較すると，表 1 のように，カバー率が向上していることが確認できる．各項目のエントロピーについても増加しており，従来のデータベースからの改善が認められた．

音声分析合成システム WORLD による統計的歌声合成を実現するため，音声符号化についても検討した[論文 2]．ここでは，サンプリング周波数が 40 kHz 以上のフルバンド音声を対象とし，符号化しない音声パラメータにより合成された結果と等価な品質が得ら得ることを条件とした符号化について検討した．この検討では，特にスペクトル包絡の圧縮について着目し，現在 1 フレームあたり 1025 次元で表現されているスペクトル包絡を，有意な品質劣化が無く 50 次元程度まで圧縮できることを示した．

最後に，スペクトルと基本周波数のモデリングを実施し，有効性を検証した[学会発表 4]．同時に，声質変換を実現するニューラルネットワークの構築を行い，有効性を示した．

表 1：歌唱データベースのカバー率

	CONV_DB	PHR_ENT_DB
音素	96.1	100.0
音高	100.0	100.0
音高差	84.8	96.9
音長	75.8	96.9
音符の小節内位置	65.6	98.9
テンポ	88.9	100.0
平均	85.2	98.8

人間性の制御を統計的歌声合成と統合する部分については現在検討中であるが，知覚モデルの構築と，声質変換を行うニューラルネットワークの組み合わせにより人間性の制御も可能である．

5．主な発表論文等

〔雑誌論文〕（計 3 件，全て査読有）

[1] M. Morise and Y. Watanabe: Sound quality comparison among high-quality vocoders by using re-synthesized speech, Acoust. Sci. & Tech., vol. 39, no. 3, pp. 263-265, 2018.
https://doi.org/10.1250/ast.39.263
[2] M. Morise, G. Miyashita, and K. Ozawa: Low-dimensional representation of spectral envelope without deterioration for full-band speech analysis/synthesis system, in Proc. INTERSPEECH 2017, pp. 409-413, 2017.
DOI: 10.21437/Interspeech.2017-67
[3] M. Morise: Harvest: A high-performance fundamental frequency estimator from speech signals, Proc. INTERSPEECH 2017, pp. 2321-2325, 2017.
DOI: 10.21437/Interspeech.2017-68

〔学会発表〕（計 12 件）

[1] 森勢将雅，豊田裕一，小澤賢司：誇張した時間的揺らぎが歌声の人間性知覚に与える影響，情報処理学会音楽情報科学研究会，2017.
[2] 本郷康貴，能勢隆，伊藤彰則：楽譜情報のバランスを考慮したフレーズ選択の検討―歌声合成のための歌唱データベースの構築に向けて―，日本音響学会 2017 年春季研究発表会，2017.
[3] F. Yokomori, M. Morise, and K. Ozawa: Effect of temporal fluctuation in speech on perception of humanness of synthesized speech, ASA-ASJ joint meeting 2016.
[4] K. Hongo, T. Nose, and A. Ito: Spectral and pitch modeling with hybrid approach to singing voice synthesis using hidden semi-Markov model and deep neural network,

ASA-ASJ joint meeting 2016.
[5] 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，2016.

〔その他〕
[1] 森勢将雅個人 Web
http://www.kki.yamanashi.ac.jp/~mmorise
[2] 音声分析合成システム WORLD
https://github.com/mmorise/World

６．研究組織
(1)研究代表者
森勢 将雅（MORISE, Masanori）
山梨大学・大学院総合研究部・准教授
研究者番号：60510013

(2)研究分担者
能勢隆（NOSE, Takashi）
東北大学・工学研究科・准教授
研究者番号：90550591