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

令和　元 年　 ６ 月 １４ 日現在

機関番号：８２６２６
研究種目：若手研究(B)
研究期間：2017～2018
課題番号：１７Ｋ１２７２１

研究課題名（和文）歌声ビッグデータを活用した歌声の多様性を考慮する歌声情報処理

研究課題名（英文）Singing information processing considering diversity of singing voice utilizing singing big data

研究代表者

中野　倫靖（Nakano, Tomoyasu）

国立研究開発法人産業技術総合研究所・情報・人間工学領域・主任研究員

研究者番号：１０５７２９２７

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：本研究では、歌声に関する大規模データセット（歌声ビッグデータ）を用いて、歌声の多様性をモデル化する要素技術開発を行った。具体的には、楽曲中の歌声分析精度向上のために、どこに歌声があるのかを推定する技術、歌詞のどの音素がいつ歌われているかを推定する技術、音高推定と歌声の分離再合成技術を、確率モデルや深層学習に基づいた手法により性能向上した。また、無伴奏の歌声のスペクトル包絡を高精度に推定する基礎技術を開発した。さらに、それらを応用するため、「何を・どう歌っているか」を同時に可視化するインタフェース、歌声の繰り返しを活用したアノテーションのための新しい歌声可視化インタフェースを実現した。

研究成果の学術的意義や社会的意義
音楽に含まれる歌声は処理が難しく未解決で本質的な課題が多い。一方で、産業・文化の両面で主要なコンテンツである音楽における最も重要な要素の一つである。したがって、学術的および産業応用的な観点からの注目度が高い。本研究の成果における歌詞同期、音高推定、歌声分離等の混合音中の歌声分析技術は、世界的に活発に研究されており、その性能向上は学術的・産業応用的に意義がある。また、そのような要素技術の性能向上が、社会・エンドユーザの音楽活動を豊かにするためには、適切なインタフェースや可視化が必要不可欠であり、その新しい技術を実現した点でも社会的に意義がある。

研究成果の概要（英文）：Fundamental technologies to model the diversity of singing voices using a large-scale data set (i.e., singing big data) were developed. Specifically, to deal with singing voice in music, methods based on probabilistic models and deep learning were developed to improve the performance of vocal activity detection, lyric synchronization, F0 (pitch) estimation, and voice separation. A fundamental technology to estimate the spectral envelope of unaccompanied singing voice with high accuracy was also developed. In order to apply those results, we realized an interface to visualize "what and how to sing" at the same time, and a new singing voice visualization interface for annotation using repetition of singing voice. Furthermore, in order to apply these methods, we implemented an interface that simultaneously visualizes "what was uttered and how the words were expressed" and a new singing voice visualization interface for annotation that utilizes repetition of singing voice.

研究分野：歌声情報処理

キーワード：歌声情報処理　信号処理　機械学習　インタフェース　情報可視化

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

(1) 本研究は、歌声に関する大規模データセット（歌声ビッグデータ）に基づいた、歌声分析・合成のための、新しい信号処理・機械学習・インタフェース技術を実現するものである。歌声ビッグデータに基づいて、歌声の多様性を編集可能性含めて考慮した研究はこれまでに存在せず、伴奏を伴う歌声を扱う点において歌声合成を視野に入れた研究は独創的である。

(2) このような歌声の多様性に関する要因を見出すことは、歌声情報処理の発展に加え、人間の歌声知覚・歌声生成の解明につながる点でも重要性が高い。

２．研究の目的

(1) 本研究は、声質や歌い方等の歌声の多様性を最大限活用し、歌声分析や合成の品質向上につなげる技術の実現を目的とする。そのために、歌声ビッグデータとしては無伴奏の歌声だけでなく、伴奏等の背景音を伴う歌声も対象に含め、より多様な声質や歌い方を活用可能とする。

(2) 背景音楽を伴う歌声を分析するための基礎技術：
楽曲中のどこに歌声があるのかを推定する技術、歌詞のどの音素がいつ歌われているかを推定する技術、歌声の音高推定技術、歌声の分離再合成技術など、背景音楽を含む楽曲中の歌声分析精度を向上させる。

(3) 背景音楽を伴う歌声を活用するための応用技術：
歌声を扱う基礎技術の性能向上を社会･エンドユーザの音楽活動の豊かさにつなげるためには、ユーザインタフェースや可視化が必要不可欠であり、多様な歌声を活用するための歌声インタフェース構築を行う。

３．研究の方法

(1) 背景音楽を含む楽曲中の歌声分析精度向上に取り組んだ。具体的には、音楽中のどこに歌声があるのかを推定する技術、歌詞のどの音素がいつ歌われているかを推定する技術、音高を推定する技術、歌声を分離する再合成技術を、確率モデルや深層学習に基づいた手法により性能向上させる。

(2) 楽曲中の歌声を高精度に分離できたことを想定し、歌声の多様性をモデル化するため、無伴奏の歌声のスペクトル包絡を高精度に推定する基礎技術の開発に取り組んだ。

(3) 上記の基礎技術やその結果を利活用するインタラクション・可視化を実現する上で、新しい歌声インタフェースの構築に取り組んだ。

４．研究成果

(1) 楽曲中のどこに歌声があるのかを推定する歌声区間推定技術（VAD）として、歌声を含まないイントロ（楽曲冒頭）の音響特徴量を機械学習することで、認識精度が向上することを確認した。また、歌声区間中どの音素がいつ歌われているかを推定する技術（歌詞アラインメント）として、音響モデル（HMM）を多様な楽曲に適応させて精度向上を確認した。
　さらに、深層学習の枠組みを用いて、音楽に含まれる歌声の音高（F0）推定と歌声信号の分離再合成技術を開発し、その性能向上を確認した。この結果は、VAD や歌詞アラインメントのさらなる性能向上にもつなげられることができる。特に、歌声分離と歌詞アラインメントに関する研究活動は、近年、世界的に活発に取り組まれており、学術的観点・産業応用的な観点からその性能向上はインパクトを持つ。また、深層学習を用いていることから、歌声ビッグデータと、そこに含まれる歌声の多様性を活用可能であり、またネットワークの組み合わせ方で様々な発展が可能となる。

(2) 無伴奏の歌声を対象に、そのスペクトル包絡を推定する新しい基礎技術を実現し、その精度向上を確認した。深層学習による音声や歌声の end-to-end 分析・合成は盛んに研究されているが、波形をそのまま扱う方法だけでなく、学習データのサンプル数を減らしたりする等の目的でスペクトル包絡が利用されることも多く、その利用価値は高い。

(3) 多様な歌声を「どう合成したいのか」に関するインタラクション（歌声インタフェース）として、歌声が「何を歌っているか」と「どう歌われているのか」を同時に可視化する技術を実現した。具体的には、発話を伴う文字テキストにおいて、各文字の発声タイミングやその音響特徴量を把握できるように可視化する TextTimeline を開発した。TextTimeline では、テキスト表示を優先しながら音響特徴を文字の周辺に埋め込むが、その際に音声の時間軸をテキストと直交する方向（横書きテキストなら縦方向）に可視化することでオリジナルの時間軸を同時に保ち、詳細な音響特徴の可視化も可能にする。
　従来、ピアノロール等のように音声の時間軸を固定（優先）して文字を分割表示する方法や、逆に、カラオケ歌詞や詩吟の吟詠譜等のように文字位置を固定して各文字の音響特徴を表現す

る方法があった。しかし前者は、文字位置が音声に制約を受け、文字間の重なりや空白によってテキストの大局的な関係を把握しにくく、後者は、音響特徴を音声の時間軸通りに可視化できない問題があった。

TextTimeline の特徴を検証するために、6 人のユーザ（男性 5 人: U1-U4, U6、女性 1 人: U5）に、以下の三種類のインタフェースを使用してもらい、定性的評価として使用コメントを得た。

D1）テキストと音声の並行表示
D2）テキストに対して音声軸を直交させた表示（テキストは横、音声軸は縦）
D3）TextTimeline

その結果、D3 は「発声タイミング」が予測しやすく、「単語間の音響特徴量の比較」がしやすいという知見を得た。ただし、描画領域が大きいという問題があるため、今後の課題である。

本研究の成果は、国内研究会で発表した（[学会発表]1）他、可視化に関する IEEE 国際会議の査読付きポスター発表を行い Honorable Mention Poster Awards を受賞した（[学会発表]2）。また、本研究成果を含む招待講演を行い（[学会発表]3）、音楽の鑑賞と創作における知的インタフェースに関する国際会議でデモを行った（[学会発表]5）。

本可視化インタフェースは、同一歌詞における歌い方の違い（歌声の多様性）を比較しながら可視化する方法として、発展可能性がある。

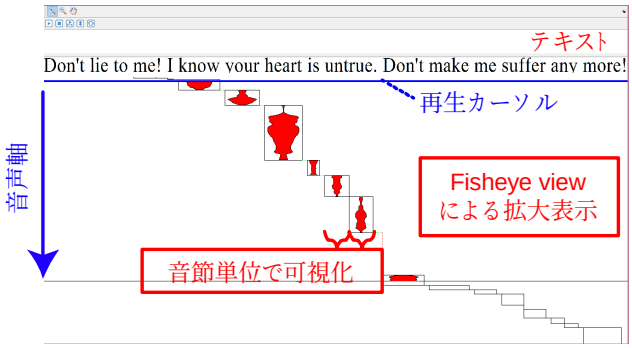

図 1 TextTimeline の画面（D3）

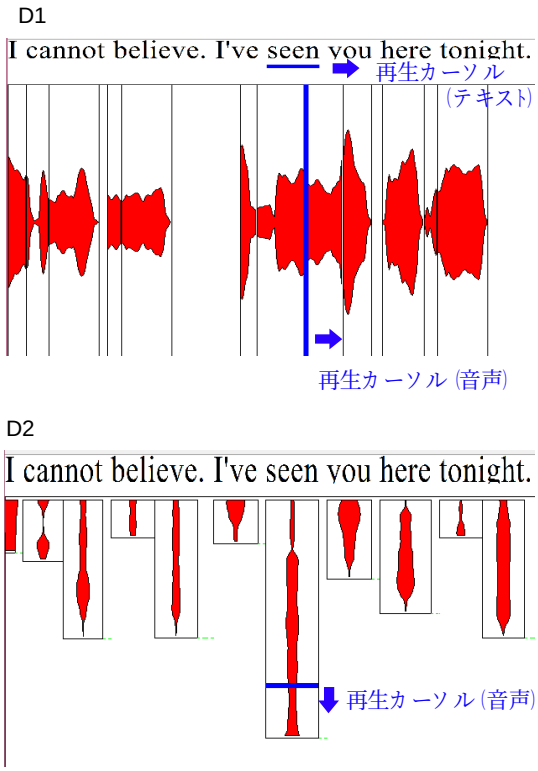

図 2 比較用の二つのデザイン（D1、D2）

(4)歌声の繰り返しを活用する新しい歌声可視化方法を実現した。ここでは、伴奏を含んだ歌声を活用する上で、音楽に含まれるボーカルの音高（F0）をアノテーションする新しいインタフェースを開発した。ソースコードエディタや表計算ソフトウェアで使用されるオートコンプリート機能のように、繰り返される類似区間を同時に可視化しながらアノテーションでき、既にアノテーションした結果を類似区間に反映させることができる。

混合音中の歌声の F0 は、歌声分離や歌唱力評価など、歌声に関する分析技術を実現する上で重要な役割を果たすため、機械学習の正解データとしてそれをアノテーションする技術が必要である。従来、混合音中の歌声の F0 アノテーションを支援する方法として、自動認識と人間による修正を可能とする半自動的なインタラクションは提案されていたが、人間の修正自体を高速化する方法はなかった。

F0 アノテーションを行った経験を持つプロの音楽家 1 名により、6 曲（2 曲：日本語男性、2 曲：日本語女性、1 曲：英語男性、1 曲：英語女性）のアノテーションをしてもらい、オートコンプリート機能のあるなしによる違いを調査した。その結果、被験者はオートコンプリート機能と類似区間の同時可視化機能について有効性があるとコメントした。特にオートコンプリート機能に関しては、類似区間が表示された場合には必ず使用しており、オートコンプリート機能がないインタフェースよりも使いやすいと述べた。また改善点としては、例えば、類似区間や自動的に推定される F0 の正しさが目で見るだけでは分からないとコメントがあったため、その確信の度合い等を可視化する等の改善点が考えられる。また、同時に可視化する類似区間は

3 つとした現状が画面領域として限界であるとコメントした一方で、表示する類似区間を切り替える機能などがあれば、3 つ以上類似区間があった場合に修正しやすいと述べた。

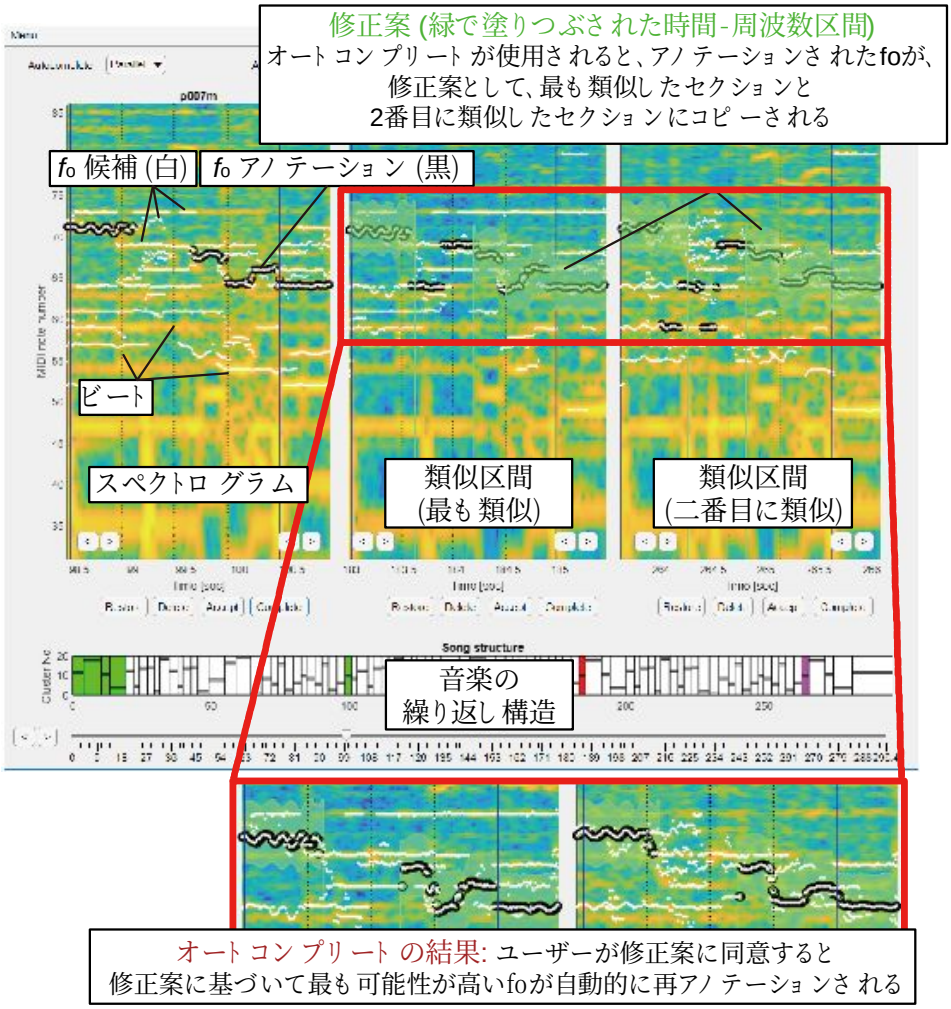

図 3 歌声 F0 のオートコンプリート・アノテーション

本研究の成果は、知的ユーザインタフェースに関する ACM 国際会議の査読付きデモ発表を行った（[学会発表]4)。

このような音楽の繰り返し構造の活用は、アノテーションだけではなく、歌声合成を用いた楽曲制作における発展につながる。例えば、基本的な歌い方を共通化させて入力したり、類似区間の歌い分けをコントロールする方法として活用できる。

５．主な発表論文等

〔雑誌論文〕（計 0 件）

〔学会発表〕（計 5 件）

1. **中野 倫靖、加藤 淳、後藤 真孝**: “TextTimeline: **文字表示を保持した発話テキストの音響特徴可視化**,” **情報処理学会 音楽情報科学研究会 研究報告**, 2017-MUS-116-21, pp.1-7 (2017)
2. Tomoyasu Nakano, Jun Kato, Masataka Goto: “TextTimeline: Visualizing Acoustic Features and Vocalized Timing along Display Text,” Proc. of the 11th IEEE Pacific Visualization Symposium (PacificVis 2018), pp.1-2 (2018)
3. **中野 倫靖**: **招待講演**: “**音楽・歌声情報処理に基づくインタフェース構築と可視化**,” **電子情報通信学会および日本音響学会 音声研究会**(SP) (2018)
4. Tomoyasu Nakano, Yuki Koyama, Masahiro Hamasaki, Masataka Goto: “Autocomplete Vocal-fo Annotation of Songs Using Musical Repetitions, ” Proc. of the 24th International Conference on Intelligent User Interfaces (ACM IUI 2019), pp.71-72 (2019)
5. Tomoyasu Nakano, Jun Kato, Masataka Goto: Demo: “TextTimeline: Visualizing

Vocalized Timing of Singing Voice along Display Text," The 2nd Workshop on Intelligent Music Interfaces for Listening and Creation (MILC 2019) (2019)

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
ホームページ等
なし

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし