歌声インタフェース：歌声を対象とした信号処理とそれに基づくインタフェース構築*

◎中野 倫靖, 後藤 真孝 (産業技術総合研究所)

1　はじめに

近年，音楽情報処理分野の発展 [1,2] と共に，歌声に関する研究活動が世界的に活発に取り組まれ [3–5]，学術的な観点からだけでなく，産業応用的な観点からも注目を集めている．そうした歌声に関する研究は，歌声固有の特徴に関する基礎研究から応用研究まで多岐に渡る．そこで我々は，文献 [3,6,7] において，そうした歌声に関する幅広い研究を「歌声情報処理」と名付け，同分野のさらなる発展を目指してきた．

ここで，歌声信号処理技術を様々なユーザに使ってもらうためには，ユーザ視点での問題発見を含め，対象ユーザの特性に合わせてインタフェース構築することが重要である．現在，プロ（レコーディングエンジニア等）向けのインタフェースは既に多く存在する（音高補正ツール Auto-tune 等）．しかし，それらは専門家向けが中心で，音楽に詳しく高度な機能を熟知していることを前提とする場合が多く，エンドユーザ向けのインタフェースは少ない．一方で，近年はエンドユーザが気軽に歌を発表したり，それを聴いて楽しんだりする文化が広がっていることから，今後はそのようなエンドユーザも歌声信号処理技術を活用できるインタフェースの構築が重要になってくる．

我々は，歌声信号処理に基づくインタフェース構築やインタラクションによって，人々の音楽生活をより豊かにする研究アプローチを「歌声インタフェース」と名付ける．歌声インタフェースという用語自体はユーザを限定するものではないが，本稿では，エンドユーザと歌との関わり合い方の未来を切り拓くことを目指した我々の歌声インタフェースの研究事例を中心に紹介する．

2　エンドユーザを対象とする重要性

近年，動画コミュニケーションサイト「ニコニコ動画（http://www.nicovideo.jp/）」等で，エンドユーザが歌を発表し（以降，歌声コンテンツと呼ぶ），それを聴いて楽しむ文化が広がっている．ここで特徴的なのは，歌声コンテンツのメインボーカルとして合成された歌声が用いられることが日常的となった点 [8,9] と，ある一つの曲（一次コンテンツ）を対象として，それを様々なユーザが自分なりに歌った二次コンテンツが数多く存在する点である [10]．

2007 年に VOCALOID2「初音ミク」[11] が発売されて以降，歌声合成ソフトウェアを用いて歌声コンテンツ制作を楽しむユーザが急増し，その利用拡大に対して技術的，社会的，文化的に注目されてきた．例えば，2013 年 7 月時点では，ニコニコ動画で「VOCALOID」を検索クエリー（タグ検索）とすると，約 27 万曲が検索結果として得られる．さらに，そうしたコンテンツを制作するユーザが増えただけでなく，それを楽しむリスナーも増えている．3200 曲以上が 10 万回，170 曲以上が 100 万回，上位 10 曲が 500 万回を超えて再生され，1000 万回以上再生された動画も存在する．

それと並行して，様々なエンドユーザが歌ったコンテンツも，Web 上で大量に公開されるようになった．これらはニコニコ動画では「歌ってみた」と呼ばれていて，投稿数や再生数の観点からは，VOCALOID による作品以上に人気がある．2013 年 7 月時点で約 63 万曲が検索結果として得られ，4650 曲以上が 10 万回，200 曲以上が 100 万回，上位 5 曲が 500 万回を超えて再生されている．

これらの投稿数の推移に関しては，「歌ってみた」に特化したものではないが，文献 [10] でも分析されている．歌声合成や「歌ってみた」に基づく作品を楽しむリスナーが増加した結果，そうしたエンドユーザによる作品が音楽 CD として市販され，その一部が商業音楽ヒットチャート上位にランクインする等 [12]，広く受け入れられてきた．

3　歌声インタフェース

歌声インタフェースの研究分野では，様々な研究者が異なる課題に取り組んでおり，大別して，歌声や音楽に関する「創作」を支援するアプローチと，「鑑賞」を支援するアプローチとがある．本章では，前者の「創作」を支援するアプローチに関する我々の研究事例を紹介し，その実現に必要な歌声信号処理技術についても述べる．

以下，3.1 では，「歌ってみた」のような自分自身の歌声によるコンテンツ制作を支援するインタフェース VocaRefiner について述べる．3.2 では，歌声合成ソフトウェアによるコンテンツ制作を支援するために，

*Singing Interface: Signal Processing for Singing Voices and Its Interface Development. by NAKANO, Tomoyasu and GOTO, Masataka (National Institute of Advanced Industrial Science and Technology (AIST))

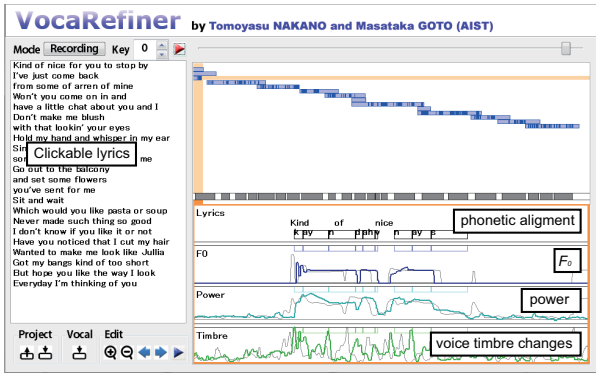

Fig. 1　VocaRefiner: 複数回の歌唱を統合できる歌声生成インタフェース [14, 15]

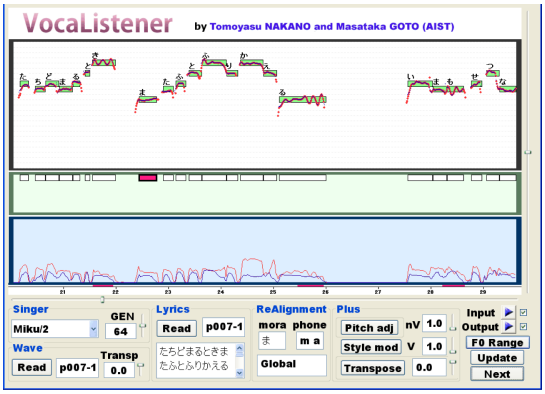

Fig. 2　VocaListener: ユーザの歌い方を真似ることが可能な歌声合成インタフェース [16]

ユーザの歌い方を真似ることが可能な歌声合成インタフェース VocaListener について述べる．3.3 では，その歌声コンテンツの魅力を上げるために，ユーザの歌唱力を向上する支援インタフェース MiruSinger について述べる．MiruSinger では，音楽 CD のボーカルの歌声や，自分自身の歌声を客観的な情報（F_0 とビブラート区間）を見ながら聴くこともできるため，歌声の「鑑賞」支援という側面も持っている．最後に，3.4 では，音楽のドラムパターンに着目して，それを歌（口ドラム）で検索しながらドラムパートを差し替えて編曲するインタフェース Voice Drummer について述べる．

その他，本稿では紙面の制約から充分に紹介できないが，後者の歌声の「鑑賞」を支援するアプローチに関しては，声質の似た歌声がメインボーカルの楽曲を探す音楽情報検索インタフェース VocalFinder [13] が挙げられる．また，ニコニコ動画上の歌声合成楽曲の視聴支援をする Web サービス Songrium [10] では，オリジナル楽曲とそれから派生した作品群を可視化し，その派生の種類によって，「歌ってみた」は「青」のように色を変えて表示する．ユーザは，「歌ってみた」の派生作品群が多ければ，そのオリジナル楽曲が歌ってみたくなる曲である，と知ることができ，一種の「鑑賞」を支援するインタフェースとしても捉えられる．

3.1　VocaRefiner: 複数回の歌唱を統合できる歌声生成インタフェース

「VocaRefiner」(図 1) は，歌声に特化して複数回の歌唱を効率的に録音するためのインタフェースであり，さらにそれらの優れた部分を統合して一つの歌声を作る歌声生成インタフェースである [14, 15]．

質の高い歌を生成するためには，ミスや乱れのない歌声が望ましいが，人間が歌う場合には，一度の歌唱で完璧に歌唱できないことが多い．そこで，何度も歌い直して録音した後，それを切り貼りして優れた部分のみを統合する処理が必要となる．しかし従来，そういった複数回歌われた歌声の扱いを視野に入れた研究はなく，複数回歌唱の効率的な録音方法，それらの効果的な統合方法を実現した．

VocaRefiner では，歌詞中の文字を指定することで，そこから歌を録音できる効率的な録音を可能とした．これによって，歌い間違えても即座に当該箇所から歌い直すことができる．そして，これらの録音における歌詞の各文字の時刻を自動推定することで，複数回の歌唱を音素単位で対応付けることができ，効率的である．さらに，それら複数回の歌唱を音の三要素（音高・音量・音色）に分解して，優れた要素を音素単位で組み替えて統合することができるようになった．

VocaRefiner では，歌声特有の特性である「歌詞」に着目することで，インタラクティブで効率的な録音を実現した点が特に新しい．

3.2　VocaListener: ユーザの歌い方を真似ることが可能な歌声合成インタフェース

「VocaListener」(図 2) は，ユーザの歌声の音高と音量を真似るように，市販の歌声合成ソフトウェアのパラメータを自動推定できる歌声合成インタフェースである [16]．従来必要だった歌声合成パラメータの長時間の調整や楽譜の入力が不要であり，お手本を歌うだけで，人間らしい自然な歌声が容易に合成できる．合成結果の具体例は，http://staff.aist.go.jp/t.nakano/VocaListener/ で視聴することができる．

本インタフェースでは，合成された歌声の音高と音量が，入力されたユーザの歌声とより近くなるように，歌声合成パラメータを繰り返し更新（反復推定）した．歌声と歌詞の高精度な自動対応付け機能を持ち，歌詞のどこをいつ歌っているかを対応付けることで，各音節の高さを推定し，音符化して歌声合成用の楽譜表現を生成可能にした．さらに，ユーザによる合

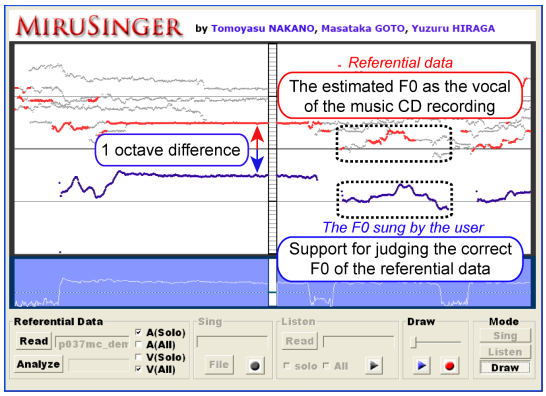

Fig. 3　MiruSinger: 歌唱力向上支援インタフェース [19]

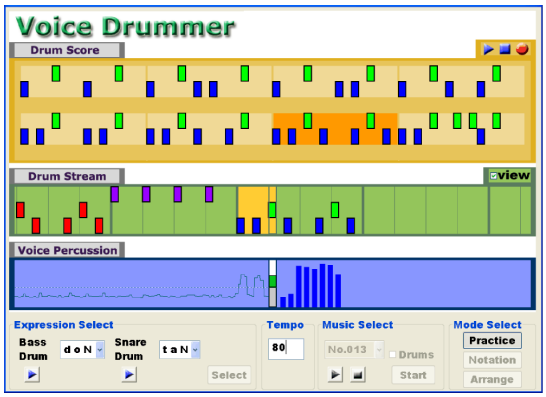

Fig. 4　Voice Drummer: 口（くち）ドラムによるドラム譜入力インタフェース [25]

成結果の微調整にも対応し，ユーザの歌唱力を補正して合成する機能も実現した [16].

関連研究としては，反復推定をせずに，ユーザの歌声から抽出した音高や音長などを直接歌声合成システムに与えた事例 [17] があったが，それに比べて合成結果を再度分析するという点が新しい．さらに，2008 年に実現された VocaListener はその後拡張され，2010 年には，ユーザの歌声の声色変化も真似る歌声合成システム「VocaListener2」[18] に発展している．

VocaListener では，入力となる歌声の質が合成結果に影響するため，自分の望む入力を VocaRefiner によって制作することも有効である．

3.3　MiruSinger: 歌唱力向上支援インタフェース

「MiruSinger」(図 3) は，既存の楽曲の歌い方をお手本として歌を練習したい時に，その楽曲のボーカルパートの F_0 を分析して可視化し，ユーザの歌声と比較表示できる歌唱力向上支援インタフェースである [19]．既存楽曲の混合音中のボーカルパートの F_0 とビブラート区間を表示して，ユーザの歌声をリアルタイムに可視化してフィードバックして比較できる．これにより，ユーザはどれぐらい自分の音高が外れているかがわかり，目標とすべき音高もわかる．

歌唱力を自動的に評価するインタフェースとしては，カラオケの採点機能が普及しており，そこでは主に評価用の楽譜情報（音高）からの差異に基づいて採点している．一方，上記の MiruSinger では，楽譜情報を用いていない点，お手本を既存楽曲中のボーカルとできる点が特長である．歌唱力向上支援インタフェースの関連研究には，ユーザ歌唱の分析結果をリアルタイムに可視化してフィードバックする事例 [20–23] や，歌唱力補正をする事例 [24] がある．

歌唱力を向上したり，歌唱の表現力を増すことを支援するこのようなインタフェースは，VocaRefiner や VocaListener を使う上でも有効である．

3.4　Voice Drummer: 口（くち）ドラムによるドラム譜入力インタフェース

「Voice Drummer」(図 4) は，ドラム音を真似た「ドンタンドドタン」のような口ドラム（ボイスパーカッション）によって，ドラム譜の入力を可能にするインタフェースである [25]．擬音語で発声された口ドラムは，事前に用意したドラムパターンのデータベースとマッチングしながら，どのパターンがどのようなテンポで歌われたかを自動的に認識・検索し，結果をリアルタイムに可視化してフィードバックする．既存の楽曲のドラムパートだけを差し替えて編曲する機能も持つ．

本インタフェースでは，ドラム音の内部表現として「ドン」「タン」のような擬音語表現を採用し，音声認識用の HMM を活用してドラムパターンデータベース中のどのパターンがユーザの口ドラムと最も一致するかを求めた点が新しい．人によって擬音語表現には個人差があるため，バスドラムとスネアドラムの口ドラムとして出現しやすい表現を調査して，認識用の辞書に登録した [25].

関連研究として，擬音語表現に近い発声や，ドラム音を音響的に模倣した音声 (Beatboxing) による音楽検索インタフェース [26, 27] 等があった．

3.5　歌声を対象とした信号処理

「歌声」を対象とした信号処理は,「話声」を扱う音声信号処理に共通する点も多い．すなわち，音高（F_0），音量，音色（スペクトル包絡，群遅延），音素時刻の推定が基本的に必要な技術である．さらに，そのような基本技術に加えて，音楽特有・歌声特有の性質を考慮して，分析やモデル化を行う場合がある．例えば，伴奏が含まれていること，メロディーや音楽的なリズムを持つこと，歌声特有の F_0 動特性や共振特性（Singer's Formant）を持つこと，歌唱様式や表情付けの違いによって，歌い方や発声のされ方に様々

なバリエーションがあることである [3, 28].

このような歌声を対象とした信号処理技術に関して我々は，口ドラム認識 [25]，歌唱力評価 [29]，ブレス検出 [30]，歌声専用音響モデルによる歌詞アラインメント [16]，声色変化に応じたスペクトル包絡の変形 [18]，スペクトル包絡推定 [31]，F_0 推定 [15]，群遅延推定 [15] を開発し，これまでに述べた歌声インタフェース構築の基盤技術として用いている．

4 おわりに

本稿では，「歌声インタフェース」と名付けた研究アプローチについて述べ，歌の創作と鑑賞を支援する様々な研究事例を紹介した．また，エンドユーザと歌声コンテンツの関わり方の現状や，歌声に関する信号処理技術についても説明した．

歌声信号処理技術が世の中で広く活用されるためには，様々なユーザを対象とした多様な歌声インタフェースが必要である．本稿で述べた歌声インタフェースにおいて，VocaRefiner では 1 度で完璧に歌えないユーザ，VocaListener や Voice Drummer では楽譜に不慣れなユーザ，MiruSinger では自分の好きな歌手の歌で練習したいユーザにとって，これまでにない方法で歌の創作や楽譜入力，歌の練習が可能になった．

今後，歌声インタフェース自体の機能を豊かで高度にしていくだけではなく，それを使うユーザ自身の歌唱力や表現力等も向上できるような研究も重要になる．歌声インタフェースを使っているうちに歌声に関するユーザの技能が向上し，新たな歌唱表現を追求したり，音楽に関する理解が深まったりできるようになると，単に技術のみが高度化していくのではない未来が切り拓ける．このように，人間と技術が相互に向上し合う未来へ向けて，歌声信号処理と歌声インタフェースを両輪として共に重視した研究開発が，様々な研究者によってより活発に取り組まれていくことが望まれる．

謝辞 本研究の一部は，JST CREST プロジェクト（CrestMuse, OngaCREST）による支援を受けた．

参考文献

[1] 後藤真孝，平田圭二：解説 “音楽情報処理の最近の研究”，日本音響学会誌，**60**，675–681 (2004).

[2] 後藤真孝：未来を切り拓く音楽情報処理，情処研報 2013-MUS-99 (2013).

[3] 後藤真孝 他：解説 “歌声情報処理の最近の研究”，日本音響学会誌，**64**，616–623 (2008).

[4] 河原英紀：歌声情報処理の最新動向，電気学会誌，**130**，360–363 (2010).

[5] 剣持秀紀：解説 “歌声合成技術の動向–「初音ミク」を支える技術–”，日本音響学会誌，**67**，46–50 (2011).

[6] M. Goto *et al.*: Singing Information Processing Based on Singing Voice Modeling, *Proc. of ICASSP 2010*, pp. 5506–5509 (2010).

[7] 後藤真孝 他：歌声情報処理: 歌声を対象とした音楽情報処理，情処研報 2010-MUS-86，No. 4, pp. 1–9 (2010).

[8] 後藤真孝：初音ミク，ニコニコ動画，ピアプロが切り拓いた CGM 現象，情報処理，**53**，466–471 (2012).

[9] 伊藤博之：初音ミク as an interface，情報処理，**53**，477–483 (2012).

[10] M. Hamasaki and M. Goto: Songrium: A Music Browsing Assistance Service Based on Visualization of Massive Open Collaboration Within Music Content Creation Community, *Proc. of WikiSym + OpenSym 2013* (2013).

[11] クリプトン：VOCALOID2 初音ミク（HATSUNE MIKU），`http://www.crypton.co.jp/mp/pages/prod/vocaloid/cv01.jsp`.

[12] H. Kenmochi: VOCALOID and Hatsune Miku Phenomenon in Japan, *Proc. of InterSinging 2010*, pp. 1–4 (2010).

[13] H. Fujihara *et al.*: A Modeling of Singing Voice Robust to Accompaniment Sounds and Its Application to Singer Identification and Vocal-Timbre-Similarity-Based Music Information Retrieval, *IEEE Trans. on Audio, Speech, and Language Processing*, **18**，638–648 (2010).

[14] 中野倫靖，後藤真孝：VocaRefiner: 歌を歌って歌い直して統合できる新しい歌声生成インタフェース，WISS 2012 講演論文集，pp. 1–6 (2012).

[15] T. Nakano and M. Goto: VocaRefiner: An Interactive Singing Recording System with Integration of Multiple Singing Recordings, *Proc. SMAC-SMC2013* (2013).

[16] 中野倫靖，後藤真孝：VocaListener：ユーザ歌唱の音高および音量を真似る歌声合成システム，情報処理学会論文誌，**52**，3853–3867 (2011).

[17] J. Janer *et al.*: Performance-Driven Control for Sample-Based Singing Voice Synthesis, *Proc. of DAFx-06*, pp. 41–44 (2006).

[18] 中野倫靖，後藤真孝：VocaListener2: ユーザ歌唱の音高・音量に加えて声色変化も真似る歌声合成システム，情処学論，**54**，1771–1783 (2013).

[19] T. Nakano *et al.*: MiruSinger: A Singing Skill Visualization Interface Using Real-Time Feedback and Music CD Recordings as Referential Data, *Proc. of ISM 2007 Workshops*, pp. 75–76 (2008).

[20] D. M. Howard and G. F. Welch: Microcomputer-based Singing Ability Assessment and Development, *Applied Acoustics*, **27**，89–102 (1989).

[21] 平井重行 他：歌の調子外れに対する治療支援システム，電子情報通信学会論文誌，**J84-D-II**，1933–1941 (2001).

[22] D. Hoppe *et al.*: Development of Real-Time Visual Feedback Assistance in Singing Training: a Review, *Journal of Computer Assisted Learning*, **22**，308–316 (2006).

[23] O. Mayor *et al.*: Performance Analysis and Scoring of the Singing Voice, *Proc. AES 35th Int'l. Conf.* (2009).

[24] K. Nakano *et al.*: Vocal Manipulation Based on Pitch Transcription and Its Application to Interactive Entertainment for Karaoke, *LNCS: Haptic and Audio Interaction Design*, Vol. 6851, pp. 52–60 (2011).

[25] 中野倫靖 他：口ドラム認識手法とそのドラム譜入力システムへの応用，情処学論，**48**，386–397 (2007).

[26] O. Gillet and G. Richard: Indexing and Querying Drum Loops Databases, *Proc. of CBMI 2005* (2005).

[27] A. Kapur *et al.*: Query-by-Beat-Boxing: Music Retrieval for the DJ, *Proc. of ISMIR 2004*, pp. 170–177 (2004).

[28] 中野倫靖：歌声分析・歌声合成，電子情報通信学会知識ベース　2 群（画像・音・言語）－ 9 編（音楽情報処理）(2010).

[29] 中野倫靖 他：楽譜情報を用いない歌唱力自動評価手法，情処学論，**48**，227–236 (2007).

[30] T. Nakano *et al.*: Analysis and Automatic Detection of Breath Sounds in Unaccompanied Singing Voice, *Proc. of ICMPC 2008*, pp. 387–390 (2008).

[31] T. Nakano and M. Goto: A Spectral Envelope Estimation Method Based on F0-Adaptive Multi-Frame Integration Analysis, *Proc. of SAPA-SCALE 2012*, pp. 11–16 (2012).