VocaListener2: ユーザ歌唱の声色変化を真似る歌声合成システム*

◎中野 倫靖, 後藤 真孝 (産総研)

1 はじめに

歌声を人工的に生成できる歌声合成システムは、多様な歌声での合成が容易に行え、歌唱の表現を再現性高くコントロールできることから、歌唱付き楽曲制作の可能性を広げる重要なツールである。我々は以前、ユーザ歌唱の音高と音量を真似るように、既存の歌声合成ソフトウェアの合成パラメータを自動調整する VocaListener を開発した [1]。これによりユーザは歌うだけで、自然な歌声を多様な音源で合成することが容易となった。しかし、音高と音量しか扱えず、ユーザ歌唱の表情や歌い方を表現しきれていなかった。そこで、本研究では声色（こわいろ）変化も歌声合成結果に反映できる VocaListener2 を提案し [2]、これまでのシステムを VocaListener1 として区別する。個人性を保持したまま、歌い方としての声色の時間変化を反映することで、表現豊かな合成を実現する。

従来、声質変換やモーフィングに関する研究がなされてきた [3, 4, 5, 6, 7] が、本研究が対象とするような、歌唱中の変化を操作することはできなかった。一方、このような声色変化を、ユーザが明示的に操作する技術には Vocaloid [8] がある。Vocaloid では、複数の数値パラメータを各時刻で調整することで、声色変化を伴った歌声合成が実現できる。しかし、曲に合わせたパラメータ操作は難しく、ほとんどのユーザはパラメータを変更しないか、変更するにしても曲毎に一括で変更したり、大まかに変更したりしていた。

2 VocaListener1 の機能と拡張

本章では VocaListener1 の概説と、VocaListener2 を実現するための課題について述べる。

2.1 VocaListener1: ユーザ歌唱の音高と音量を真似る歌声合成パラメータ推定システム [1]

VocaListener1 は、既存の歌声合成ソフトウェアの歌声合成パラメータを、ユーザ歌唱からその音高と音量を真似て推定する技術である（図 1）。パラメータの反復推定により、推定精度が従来研究 [9] に比べて向上し、歌声合成システムやその音源（歌手の声）を切り替えても再調整せずに自動的に合成できる [1]。また、独自の歌声専用音響モデルによって、歌詞のテキストを音符毎に割り当てる作業はほぼ自動で行える。音符の割り当てでは、その推定時刻に誤りが発生

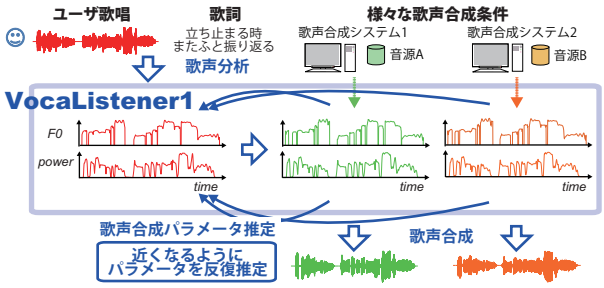

図 1 VocaListener1 による歌声合成パラメータ推定。

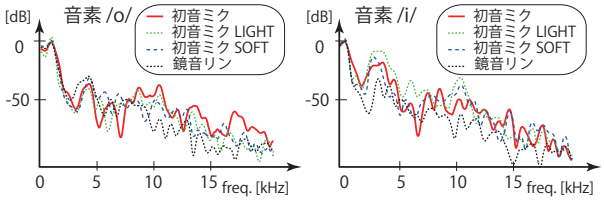

図 2 音韻の違いと歌唱者の違いによるスペクトル包絡形状の違いの具体例。

する可能性があるが、誤った箇所を指摘（ダメ出し）するだけで、新しい候補を再提示する機能もある。

2.2 声色変化を真似る歌声合成への拡張

声色変化を対象として「ユーザ歌唱を真似る」ためには、VocaListener1 と同様、既存の歌声合成システムにおけるパラメータをユーザ歌唱に合わせて自動推定する方法が考えられる。しかし、声質や声色変化に関するパラメータは、システム毎に異なる可能性が高く、実際、Vocaloid と Vocaloid2 [8] では一部異なる。また、Vocaloid2 の応用商品である「初音ミク・アペンド [10]」は、「初音ミク [11]」と同一歌唱者の声で、DARK, LIGHT, SOFT, SOLID, SWEET, VIVID の 6 種類の声色で歌声合成できる。しかし、これらの音源をフレーズ毎に切り替えながらの合成はできても、歌声合成システム上でこれらの中間の状態を作り出すことは困難である。

したがって、システムやパラメータ毎に最適化した方法を仮に実現しても、異なるシステムにおいて適用できず、汎用的でない。つまり、歌声合成システム内のパラメータ操作だけでは不十分で、外部の信号処理が必要といえる。そのような方針を採ることで、複数音源を活用する等、自由度を拡げることにも繋がる。その上で、以下の二つの課題を解決する必要がある。

[課題 1] 声色変化をどう表現するのか

[課題 2] ユーザの声色変化をどう反映させるのか

---

*VocaListener2: A Singing Synthesis System Mimicking Voice Timbre Changes of User's Singing. by Tomoyasu NAKANO and Masataka GOTO (AIST)

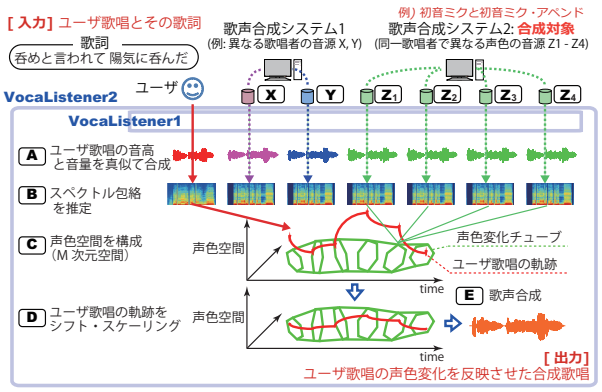

図 3　VocaListener2 における処理の流れ。

本研究では、声色の違いを初音ミクと初音ミク・アペンドの違いに相当すると考え、音響的にはスペクトル包絡の形状の違いとして定義する（図 2）。ただし、スペクトル包絡には音韻性や個人性も含まれるため、そういった成分を抑制した時間変化が声色変化といえる。このようにして、声色変化を反映したスペクトル包絡系列を生成することで本システムを実現する。

3　VocaListener2: ユーザ歌唱の声色変化を真似る歌声合成システム

2.2 節で述べた課題 (1) を解決するために、ユーザ歌唱を真似て、各時刻において音高・音量・音韻が同期した、複数の歌唱者による歌唱音声を自動生成する。これらのスペクトル包絡から、声質や声色変化に寄与する成分以外を部分空間法に基づいた処理によって抑制して**声色空間** (M 次元) を構成する。このような方法は、音韻性と話者性の分離に基づく話者認識 [12] や声質変換 [13] で有効性が確認されている。

ここで、声色空間の構成には、合成対象となる複数の声色の歌唱も用いることで、各時刻でそれぞれが空間上の一点として表現され、その時間変化として複数の軌跡が得られる。この、各声色の点を含む M 次元の多面体とその時間軌跡を**声色変化チューブ**と呼ぶ。すなわち、この時々刻々と変化する多面体に囲まれた内側が声色変化可能な領域と本研究では仮定する。

続いて、課題 (2) を解決するために、同じ声色空間の別の場所に存在するユーザ歌唱の軌跡を、声色変化チューブ内になるべく入るように位置やスケールを合わせる。これによって、各時刻における声色空間上の合成目標位置が決定される。最後に、その位置からスペクトル包絡を生成できれば、ユーザ歌唱の声色変化を真似る歌声合成を実現できる。

3.1　VocaListener2 の処理概要

処理の流れを図 3 に示す。まず、入力としてユーザの歌唱音声を与え、VocaListener1 を用いて、その

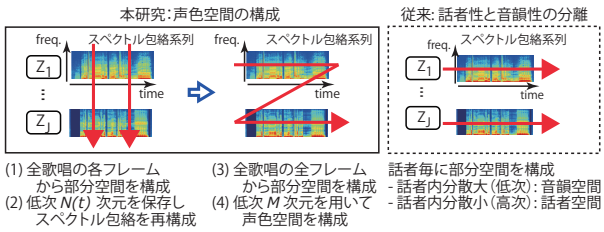

図 4　声色空間の構築と従来研究 [12, 13] との違い

歌い方を真似た歌声を複数生成する（A）。続いて、それぞれの歌唱音声からスペクトル包絡を推定して（B）、それらから M 次元の声色空間を構成する（C）。最後に、声色空間上のユーザ歌唱の軌跡を Z_1〜Z_4（合成対象となる声色：初音ミクと初音ミク・アペンドに相当）によって構成される声色変化チューブによく収まるように、シフトとスケーリング操作を行ない（D）、ユーザ歌唱の声色変化を反映して歌声合成する（E）。以降、それぞれの処理について、本稿における具体的な実現方法を説明する。歌唱音声はサンプリング周波数 44.1kHz のモノラル信号を扱い、処理の時間単位は 1 ms とする。

3.2　歌声分析

スペクトル包絡の推定と声色空間の構成、ユーザ歌唱の軌跡の調整について述べる。

スペクトル包絡の推定 B: STRAIGHT [14] を用いて推定する。分析フレーム長は F_0 同期とし、各時刻でスペクトル包絡を推定する際の FFT 長は 4096 点とした。また、離散コサイン変換を行って次元数を落とし、0 次（直流）を除いた低次 80 次元を用いた。

声色空間の構築 C: 図 4 に処理の概要を示す。まず全歌唱の各分析フレーム（時刻）で主成分分析を行う。時刻 t において、その低次 $N(t)$ 次元のみを保存して元の空間に戻すことで、各フレームの共通成分（音韻性など）を抑制する。次元数 $N(t)$ は累積寄与率 80%を超えるようにフレーム毎に決定した。最後に、全歌唱の全フレームを用いて一度に主成分分析を行い、その低次 $M(=3)$ 次元の空間を声色空間として扱う。このような処理によって、異なる歌唱者の全てのフレームが同じ空間上で扱えるだけでなく、音韻などの文脈に伴なう声色変化に関係する成分を、低次元で効率的に表現できる。これらの処理は、全ての歌唱で F_0 が存在する有声区間のみを用いた。

ユーザ歌唱の軌跡の調整 D: ユーザ歌唱の軌跡と声色変化チューブのそれぞれが平均が 0、標準偏差が 1 となるように、次元毎にシフト・スケーリングする。

3.3　歌声合成

声色空間上の一点から、それに対応付くようなスペクトル包絡を生成する E。ここで、声色空間を構成

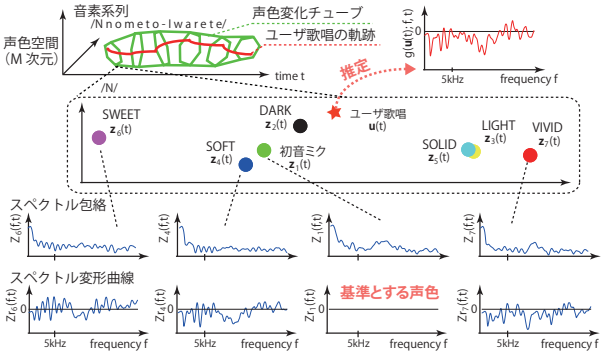

図 5　実際に構成した声色空間上のある時刻における初音ミクと初音ミク・アペンド（DARK, LIGHT, SOFT, SOLID, SWEET, VIVID）の配置と、それに対応するスペクトル包絡とスペクトル変形曲線。また、実際の推定結果。

するために利用する歌声合成システムには、Vocaloid と Vocaloid2 [8] を採用し、その応用商品として市販されている歌声合成ソフトウェアのうち、日本語歌唱を合成できる 17 種類[1]を用いた。そのうち、初音ミクと初音ミク・アペンド（DARK, LIGHT, SOFT, SOLID, SWEET, VIVID）の 7 種類を合成対象の歌声とし、それらから声色変化チューブを構成する。

ある時刻における実際の声色空間上での配置を図 5 に示す（ただし、左上の声色変化チューブはイメージ図である）。それぞれの点にはスペクトル包絡が対応付いており、これに基づいて、ユーザ歌唱の声色変化を反映させるスペクトル包絡を生成する。ただし、スペクトル包絡は直接推定せず、標準的な声[2]を基準としてそこからの変形比率をまずフレーム毎に推定する。本稿では、これを**スペクトル変形曲線**と呼び、全時刻のスペクトル変形曲線を合わせて**スペクトル変形曲面**と呼ぶ。時刻 t、周波数 f におけるスペクトル変形曲面 $Zr_j(f,t)$ は、以下のように定義する。

$$Zr_j(f,t) = \log\left(\frac{Z_j(f,t)}{Z_1(f,t)}\right) \tag{1}$$

ここで $Z_{j=1,2,\cdots,J}(f,t)$ は、j 番目の声色のスペクトル包絡を表す[3]($j=1$ が基準の声色)。$Zr_j(f,t)$ は、対数をとることで結果が負の値を取ることを許容する。続いて、声色空間上でのユーザ歌唱を $\mathbf{u}(t)$、各声色を $\mathbf{z}_j(t)$ とし、声色空間上でのユーザの声色を真似るためのスペクトル変形曲線 $g(\mathbf{u}(t);f,t)$ を推定する。そのために、Radial Basis Function を用いた Variational Interpolation [15] を応用して適用する。

$$g(\mathbf{u}(t);f,t) = \sum_{k=1}^{J}\left(w_k(f,t)\cdot\phi\left(\mathbf{u}(t)-\mathbf{z}_k(t)\right)\right) + P(\mathbf{u}(t);f,t) \tag{2}$$

$$Zr_j(f,t) = \sum_{k=1}^{J}\left(w_k(f,t)\cdot\phi\left(\mathbf{z}_j(t)-\mathbf{z}_k(t)\right)\right) + P(\mathbf{z}_j(t);f,t) \tag{3}$$

[1]男性歌唱が 3 種類、女性歌唱が 14 種類。
[2]例えば初音ミク・アペンドでない初音ミク。
[3]本稿では $J=7$ である (初音ミクと初音ミク・アペンド 6 種)。

$$g(\mathbf{z}_j(t);f,t) = Zr_j(f,t) \tag{4}$$

$$P(\mathbf{x};f,t) = p_0(f,t) + \sum_{m=1}^{M} p_m(f,t)\cdot x^{(m)} \tag{5}$$

ここで、w_j が混合比率、$P(\cdot)$ は式 (5) のように、ベクトル $\mathbf{x}$ として $\mathbf{z}_j(t)$ もしくは $\mathbf{u}(t)$ を変数とする M 変数一次多項式（係数が $p_{m=0,\ldots,M}$）である。$\phi(\cdot)$ は、ベクトル間の距離を表す関数であり、本稿では $\phi(\cdot)=|\cdot|$ とする。上記の式により、ユーザ歌唱が声色空間上で各声色の点と重なりあった場合には、それと同じスペクトル変形曲線を生成する制約（式 (4)）を満たすよう推定することになる。声色空間を $M=3$ 次元とすると、時刻 t 毎に以下の行列として書ける。

$$\begin{bmatrix}
\phi_{11} & \cdots & \phi_{1J} & 1 & z_1^{(1)} & z_1^{(2)} & z_1^{(3)} \\
\phi_{21} & \cdots & \phi_{2J} & 1 & z_2^{(1)} & z_2^{(2)} & z_2^{(3)} \\
\vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\
\phi_{J1} & \cdots & \phi_{JJ} & 1 & z_J^{(1)} & z_J^{(2)} & z_J^{(3)} \\
1 & \cdots & 1 & 0 & 0 & 0 & 0 \\
z_1^{(1)} & \cdots & z_J^{(1)} & 0 & 0 & 0 & 0 \\
z_1^{(2)} & \cdots & z_J^{(2)} & 0 & 0 & 0 & 0 \\
z_1^{(3)} & \cdots & z_J^{(3)} & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_J \\ p_0 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix}
=
\begin{bmatrix} Zr_1 \\ Zr_2 \\ \vdots \\ Zr_J \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{6}$$

ここで ϕ_{ij} は $\phi(\mathbf{z}_i(t)-\mathbf{z}_j(t))$ を表し、(f,t) や (t) は省略して記述した。

このようにして推定された w_j と p_m を用いて、式 (2) によってスペクトル変形曲面を生成する。続いて、合成の不自然さを減らすために、フレーム毎に上限と下限を設けた閾値処理を行い、時間–周波数平面上の平滑化処理により、急峻すぎる変化を低減してスペクトルの連続性を保つ。最後に、基準とした歌唱音声のスペクトル包絡を、このスペクトル変形曲面を用いて変形し、それを STRAIGHT で合成することでユーザ歌唱の声色変化を真似た合成歌唱を得る。

3.4　インタフェース構築: 声色変化の調整機能

ユーザ歌唱を真似るだけでは、歌唱によるユーザの表現力の限界を超えることができないため、以下の三つの機能を持つインタフェースを提案する。

スケーリング機能: 声色変化のスケールを変えて抑揚をつけたり、逆に抑制して歌声を合成できる。

シフト機能: 声色変化の中心を変えることで、それぞれの声色を中心とした声色変化に変換できる。

シフト・スケーリング機能: 上記二つの機能を部分的に適用することで、細かな修正を可能とする。

4　実験

本システムの評価として、RWC 研究用音楽データベース（音楽ジャンル）[16] RWC-MDB-G-2001 No.91「大漁船」を用いて 2 種類の実験を行った。本実験では、便宜上のユーザ歌唱として「大漁船」の無伴奏の男性

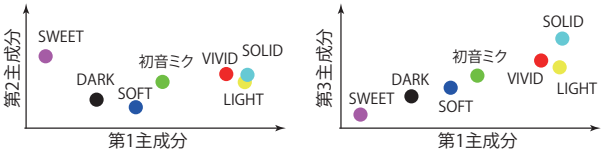

図 6　声色空間の低次 3 次元における「大漁船」の声色毎の全フレーム（55 sec）における平均ベクトル。

歌唱（55 秒）を用い、声色空間の構成では、VocaListener1 での合成歌唱が女性の場合、VocaListener-plus [1] によって 1 オクターブ上げて合成した。

4.1　実験 A: 声色空間の特性の確認

部分空間法による声色空間の特性を確認する。図 6 に、構成された $M(=3)$ 次元の声色空間とその平均ベクトルを示す。フレーム毎に 17 種類の歌唱音声に対して行った主成分分析では、$N(t)$ の平均は 4.18 次元であり、全歌唱の全フレームにおける主成分分析では、上位 3 次元の累積寄与率は 42.3%であった。

図 6 の結果から、声色空間上で各声色の平均ベクトルは比較的分離しており、また定性的ではあるが、初音ミクと初音ミク・アペンドのそれぞれの合成歌唱の聴取印象を反映するような配置となっていた。以上の結果から、低次部分空間を活用することで、声色空間を少ない次元で効率的に表現できるといえる。

4.2　実験 B: 声色変化の推定結果の確認

「大漁船」の男性歌唱を初音ミク・アペンドで真似た六種類の歌唱を合成し、それを手作業で切り貼りした歌唱音声を入力とする。このような実験によって、声色変化の推定結果を確認する。実験では、声質に関するパラメータ（GEN）をデフォルトとした場合（Closed 実験）と、GEN を 90 に変更して 2 半音下げた合成歌唱（Open 実験）の二種類を入力とした。

図 7 に、声色空間上におけるそれぞれの声色とのユークリッド距離を示す。Closed 実験の結果からは、正解がほぼ推定できたことから、声色空間の構成と、ユーザ歌唱軌跡のシフト・スケーリングが適切に行えているといえる。また Open 実験の結果では、図 6 で近くに配置された声色が相互に影響を受けていることがあった。ただし、全ての声色との距離からスペクトル包絡を推定する（式 2）ため、最近傍の声色が正解と異なっても近くに配置されていさえすれば、適切な包絡が推定できる。以上の結果から、ユーザの声色変化を真似るための有効性が示唆された。

5　おわりに

本稿では、ユーザ歌唱からの声色変化の推定と、それを真似て歌声合成する VocaListener2 を提案した。合成結果の具体例は http://staff.aist.go.jp/t.nakano/VocaListener2/ で視聴できる。声色変化は音高や音量と違い、物理量として単純に扱うことができず、未解決な課題も多い。今後は声色変化をモデル化して再利用する等、声色変化の新たな活用法について更なる検討をしていきたい。例えば、初音ミクを基準としたスペクトル包絡の変形曲面を別の音源に適用することで**声色転写**が行える可能性がある。実際、初音ミクから初音ミク・アペンドへの 6 種類の変形曲面を、そのまま鏡音リンに適用し、6 種類の「鏡音リン・アペンド」に相当する印象が得られたことを定性的に確認した。しかし、このままでは単純な処理で、汎用的でないため、今後の検討が必要である。

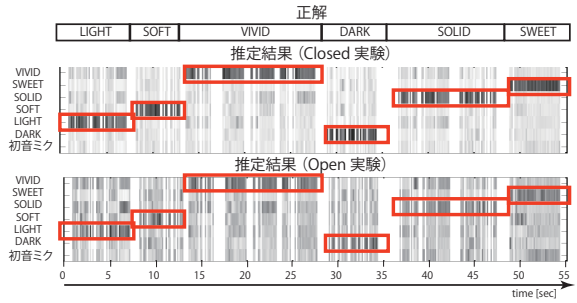

図 7　ユーザ歌唱として初音ミク・アペンド歌唱を切り貼り（LIGHT, DARK, SOLID, SOFT, VIVID, SWEET の順）した歌唱を与えた場合の声色空間上での各声色とのユークリッド距離（濃いほど距離が近い）。上段は声色空間を構成する際の初音ミク・アペンドと同じ歌唱を入力として与え、下段は GEN パラメータを変更（GEN=90）して 2 半音下げた歌唱を入力として与えた場合。

本研究の根底には、「人間らしい歌唱」とは何かを解明し、より人間を知ることがあり、本システムは、そうした歌声研究の基本ツールとしても貢献できる。

謝辞　本研究の一部は、CrestMuse プロジェクトによる支援を受けた。また本研究では、RWC 研究用音楽データベース (音楽ジャンル RWC-MDB-G-2001) を使用した。

参 考 文 献

[1] T. Nakano *et al.*: VocaLis ..., *SMC 2009*, 343–348, 2009.
[2] 中野 他: VocaListene..., 情処研報, 2010-MUS-86, 3, 2010.
[3] T. Toda *et al.*: Voice conversion based on maximum likelihood ..., *IEEE Trans. ASLP*, **15**, 2222–2235, 2007.
[4] 大谷 他: STRAIGH ..., 信学論, **J91-D**, 1082–1091, 2008.
[5] O. Türk *et al.*: A comparison of voice conversion methods for transform ..., *Interspeech 2008*, 2282–2285, 2008.
[6] T. Nose *et al.*: HMM-based style control for expressive ..., *IEICE Trans. Inf. & Syst.*, **E92-D**, 489–497, 2009.
[7] Z. Inanoglua *et al.*: Data-driven emotion conversion in sp ..., *Speech Communication*, **51, Is. 3**, 268–283, 2009.
[8] H. Kenmochi *et al.*: VOCALOID – Commercial Singing Synthesizer base ..., *Interspeech 2007*, 4011–4010, 2007.
[9] J. Janer *et al.*: Performance- ..., *DAFx-06*, 41–44, 2006.
[10] http://www.crypton.co.jp/cv01a/
[11] http://www.crypton.co.jp/cv01/
[12] 西田 他: 音韻性を抑 ..., 信学論, **J85-D2**, 554–562, 2002.
[13] 井上 他: 部分空間と ..., 信学技報 SP, 101, 86, 1–6, 2001.
[14] H. Kawahara *et al.*: Restructuring speech representations usi ..., *Speech Communication*, **27**, 187–207, 1999.
[15] G. Turk *et al.*: Modelling with implicit surfaces that interpolate, *ACM Trans. on Graphics*, **21**, 855–873, 2002.
[16] 後藤 他: RWC 研究用音 ..., 情処学論, **45**, 728–738, 2004.