3 歌声合成とその応用

剣持 秀紀
ヤマハ ST 開発センター

歌声合成の盛り上がり

最近，歌声合成技術が注目を集めている．動画投稿サイト「ニコニコ動画」には，「初音ミク」（図 -1）を筆頭とする歌声合成ソフトウェア Vocaloid を用いて合成した音声を使った動画が溢れ，市井のクリエイタたちが日夜新曲を競って発表している．クリエイタが合成音声を用いて制作したオリジナル楽曲に対して，別のユーザが動画を付けたり，アレンジを変えた新たな楽曲を制作したりするという，今までの楽曲制作にないスタイルでの協業も行われている．合成音声によるオリジナル楽曲を別のユーザが自分の声で歌って投稿するという，いわば人間と機械の主客逆転の現象も起きていることも興味深い．

合成音声による楽曲の人気はネットの中にとどまらない．動画投稿サイトで人気が出た楽曲はカラオケや着うたとして配信されたり，いわゆるメジャーレーベルから CD として発売されたりしている．2009 年 3 月 4 日にソニー・ミュージックエンタテインメントから発売されたアルバム "supercell" は，ボーカルパートはすべて「初音ミク」による合成音声であるが，オリコンによる発売日当日の売り上げランキングで 2 位を記録している．このほかにも Vocaloid を用いて制作した楽曲による CD が何枚も発売されている．

「初音ミク」は大ヒットとなり，音楽制作用のソフトウェアとしては異例の 4 万本以上の売り上げとなっている．

このように歌声合成はかつてないほど注目されている．本稿では，歌声合成の意義についての考え，歌声合成の歴史について触れた後，筆者が開発に携わった Vocaloid 歌声合成システムを紹介し，そして最後に歌声合成の今後の方向性について述べる．筆者は一企業の研究開発部門に在籍する身であるので，本稿では純粋な技術解説だけでなく，歌声合成のビジネス面に関する議論も含まれることをあらかじめご了承いただきたい．

図 -1 「初音ミク」

歌声合成の意義

■ なぜ歌声合成が必要か？

さて，そもそもなぜ歌声の合成が必要なのであろうか．人間が歌えば済む話なのに，なぜわざわざ合成音声を使用する必要があるのだろうか．これをピアノと電子ピアノの関係に置き換えて考えてみる．電子ピアノの購入者はなぜピアノではなく電子ピアノを購入するのだろうか．それは，電子ピアノの購入者にとって電子ピアノはピアノの単なる代用品ではなく，代用品以上の価値があるからである．すなわち，ピアノに比べて電子ピアノの方が持ち運びしやすい，調律の必要もない，色々な音色や機能を持っている等の点が単なる代用品以上の価値となっているわけである．歌声合成も，人間歌唱の単なる代用品としてだけでなく，合成音声でしかできないことがなければ世の中には受け入れられないだろう．

筆者が Vocaloid を開発していた頃や発売後間もない頃「何時間もかけて合成音声を作り込むくらいなら，歌

手を呼んできたほうが安いし早い」と言われたものである．しかし，今にして考えてみれば，この批判は歌声合成が受け入れられる条件を逆に示していたとも言える．すなわち，歌手を呼んできても実現できないことが可能，または歌手を呼んでくるよりも「安くて早い」のであれば，作り込みに多少の時間がかかっても受け入れられるということである．つまり「初音ミク」のようないわゆる「かわいい声」で，正確な音程で，長時間不平も言わずに歌ってくれる歌手を探すのは難しい．また，最近では，動画サイトで「初音ミク」を使用した楽曲を投稿すると，他のユーザに聴いてもらえるからという理由で曲作りに「初音ミク」を使用するクリエイタも存在する．これらは実在の歌手を呼んできて歌ってもらったとしても実現できないことである．代用品ではなく，歌声合成でなければ実現できないことを訴求できなければ，歌声合成の存在意義はないと言えるだろう．

■合成から見た歌声～楽音と音声という2つの側面

歌声には楽音と音声という2つの側面がある．まず楽音としての側面を考えてみる．楽器音と歌声の最大の違いは，歌声には歌詞がある，という点である．これはさまざまな異なる楽器をリアルタイムに順次切り替えながら演奏しているということに等しい．さらに注意すべきは楽器の場合と異なり，音の出だしのタイミングと音符のタイミングが異なる場合があるという点である．すなわち，ある音符に割り当てられている歌詞が子音＋母音という構成の音節の場合，その音符のタイミングは子音開始の位置ではなく母音開始の位置になるということである．これは合成という観点から見ると楽器音とは異なる取り扱いが必要となる．楽器音の場合は，音符開始位置（つまりMIDIでのNote ON）の位置で発音開始とすればよいが，歌声の場合は音符開始位置より前に発音を開始しなければならない．いわば因果律に反するようなことを行わなければならない．伴奏との同期を考えると，合成においてはこのことは無視できない．

次に音声としての側面を考えてみる．歌声と歌声以外の音声との違いは，歌声では音程とタイミングが楽譜（あるいはそれに相当するもの）によりある程度支配されているという点である．これにより歌声の韻律はそれ以外の音声の韻律と比べて著しく異なったものになっている．また音符の長さや組合せにより，韻律は自由に変化するので，歌詞との組合せを考えると，テキスト音声合成で行われているような大規模コーパスベースの素片連結による合成システムは事実上不可能だといえる．

さらに歌声で注意すべきは，声そのものが審美的な対象になるということである．合成という観点から見ると，特に伸ばし音の「美しさ」は歌声合成の場合必須となる．また，合成された歌声が楽曲の中で使用され，鑑賞の対象になるということから，合成音の品質はいわゆるハイファイであることが求められ，少しのノイズであっても許容されない．素片接続により合成するシステムの場合，接続境界でのノイズをいかに減らすかという点が重要になってくる．

■歌声合成に求められる要件

以上を踏まえ，実際のアプリケーションとしての観点から，歌声合成に求められる要件について考えてみたい．筆者は，歌声合成システムに求められる要件として，(1)了解性，(2)自然性，(3)操作性の3つを考えている．

(1) 了解性

スキャット的な歌唱を除いて，大抵の歌声には歌詞が伴う．合成された歌声の歌詞が聞き取れるということは，歌声合成システムの最低限の条件であると言える．

(2) 自然性

合成された歌声は人間の歌声に近い「自然」な音声でなければならない．人間の歌声に含まれるピッチの自然な揺らぎや息の成分ができるだけ再現されていることが望ましい．

(3) 操作性

システム全体としての操作性，使いやすさも重要な条件である．また，合成音を単独で使用することは少なく，伴奏音と組み合わせたり，合成音自体にもコンプレッサやリバーブ等のエフェクトをかけて使用することから，既存の音楽制作環境との連携性も重要である．

歌声合成の歴史

さて，ここでこれまでの歌声合成の歴史を簡単に振り返ってみたい．

■歌声合成の研究

1962年にベル研究所のKellyらによって発表された歌声合成は，世界初の歌声合成とされている．そのときにMax Mathewによって作られた "Daisy, daisy ..." という歌声は，文化的にも大きな影響を残し，1968年に公開された映画「2001年宇宙の旅」の最後のシーンでコンピュータHAL9000が停止する直前に "Daisy, daisy ..." と歌う場面にも影響を与えたと言われている．Kellyらの音声合成は，音響管モデル（acoustic tube model）と呼ばれるもので，滑らかに管の直径が変化するという簡単な形で声道を表現したものである[1)]．

物理モデルによる歌声合成としては，1992年にPerry Cookによって発表されたSPASMというシステムが知

られている．これはより精緻なモデルで表現したものであり，鼻道などの表現もできるようになっている．

物理モデルによる合成は，パラメータと物理量が直結していて分かりやすいという長所はあるが，精密にモデリングしようとすればするほど扱うパラメータの数が膨大になるという欠点もある．

さて，歌声も音声の一種であるので，音声の研究の成果の多くも歌声合成に活かされている．線形予測符号化（LPC）およびそれから導かれるソースフィルタモデルの歌声合成への貢献も計り知れない．1980 年に Klatt らが発表した MITalk（のちの DECTalk）は，2 次 IIR フィルタ群の並列および直列構成を使用している．DECTalk による歌声合成もよく知られている．ストックホルム王立工科大学（KTH）では，伝統的に歌声合成や歌声のモデリングに関する研究が盛んに行われているが，KTH でも同様のフォルマントモデルを歌声合成に応用した MUSEE DIG と呼ばれるシステムが知られている．

物理モデルもソースフィルタモデルも音声の生成をモデリングしたものとなっているが，一方では音声の生成過程にとらわれず，発音された音のスペクトルそのものをモデリングする手法も数多く歌声合成に取り入れられている．McAulay らによって発表された正弦波モデリング [2] は，音楽分野でも多くの応用を生み出した．正弦波モデリングでは音声信号の短時間 FFT により正弦波の強度，周波数および位相を時間的に変化する関数として求める．正弦波モデリングによる歌声合成も提案されている．

音声のスペクトルをモデリングする手法の一種として，IRCAM の CHANT というシステムに使用されている時間領域のフォルマント波形関数（Formant wave function）による手法もよく知られている [3]．これはフォルマント 1 つ 1 つのインパルス応答を時間領域の波形として表現し，その重ね合わせにより歌声を生成するものである．

■ 商用歌声合成システム

研究レベルの歌声合成システムだけでなく，これまでにいくつかの商用の歌声合成システムも市販またはサービス提供されている．そのうち代表的なものを紹介したい．

1997 年にヤマハから発売された PLG-100SG というプラグインボードは，ヤマハの MIDI 音源に組み込んで使用する機能拡張ボードであり，歌声を合成することが可能になっている．合成方式は FM 音源をベースとした時間領域フォルマント波形合成方式である．

1999 年に発売された KAE Labs（カナダ）による Vocal Writer は，Macintosh 用の歌声合成ソフトウェアである（合成方法は不明）．歌声だけでなく，伴奏のパートもこのソフトウェア上で生成可能となっている．

2000 年に NTT より発表された HORN 法は正弦波重畳により歌声を合成する手法であるが，この方式を利用したワンダーホルンという歌声合成ソフトウェアが NTT アドバンストテクノロジーより発売されている．

2004 年に Virsyn（ドイツ）から発売された CANTOR というソフトウェアは，音楽制作環境に特化したインタフェースを持ち，歌声を合成することができる（合成方式は不明）．

ヤマハが 2003 年に発表し，2004 年に最初の製品が発売された Vocaloid は，商品としてはヤマハとライセンス契約を結んだ各社が独自に制作した歌手ライブラリに，ヤマハが開発したソフトウェアが同梱される形で，ライセンス供与先の製品として発売されている．歌手ライブラリの違いにより，別の製品という扱いで発売されている．2009 年 7 月時点で，12 種類の製品が発売されている．次章では弊社が開発に携わったこの Vocaloid 歌声合成システムの技術的内容について紹介する．

そのほかにも市販のテキスト音声合成ソフトウェアで，合成器に歌わせる機能を持たせたものもいくつか発売されている．

ゲームの分野では，歌声合成の機能が組み込まれたプレイステーション 2 用のゲーム「くまうた」（2003 年 ソニー・コンピュータエンタテインメント）は，作成した歌を熊に歌わせるという独特の世界観が注目され，現在でも根強い人気を持つ．

Vocaloid 歌声合成システム技術紹介

Vocaloid は素片連結型の歌声合成システムである．歌手の歌声から取り出した音声素片を入力された楽譜情報に合うように接続することで合成を行っている．図 -2 に示されるように，(a) スコアエディタで入力された歌詞と音符を (c) 合成エンジンが受け取り，(b) 歌手ライブラリから適切な素片を選択し，接続，合成を行う．

以下，それぞれの構成要素について述べる．

(a) スコアエディタ

歌声合成のための入力インタフェースでは，画面上で歌詞と音符の対応関係が分かるように表示される必要がある．Vocaloid の入力インタフェースでは，音符はピアノロールで入力し，音符の上に直接歌詞を入力できるようになっている．

現状では日本語と英語の歌詞を入力可能である．日本語の場合は仮名またはローマ字で，英語の場合は単語そのものを入力する．入力された歌詞は自動的に音素列に変換される．

ビブラートなどの表情は，音符付近に表示されるアイ

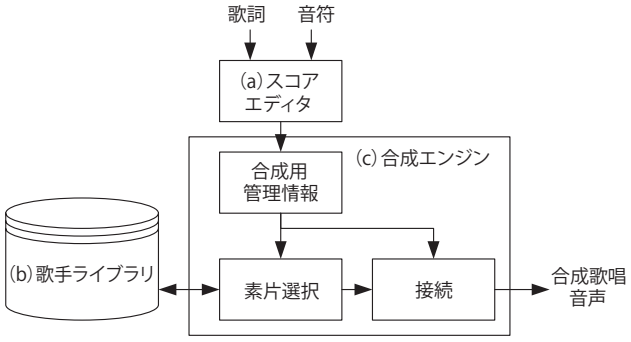

図 -2　Vocaloid の構成

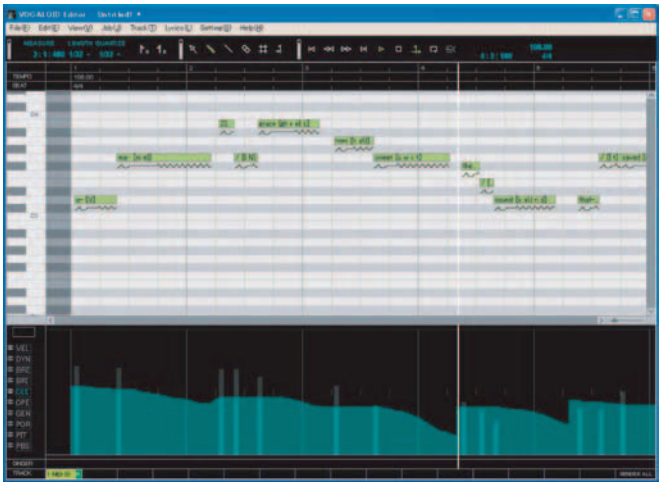

図 -3　スコアエディタ

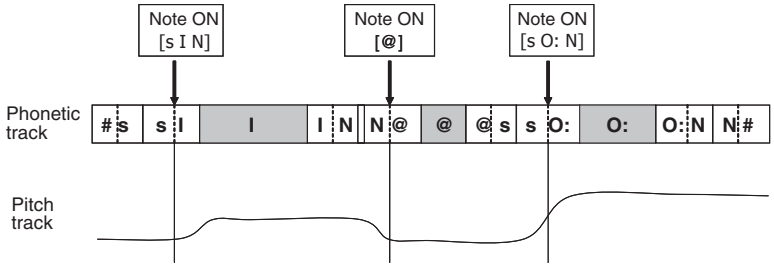

図 -4　素片使用のタイミング調整とピッチカーブ
(“Sing a song” [sIN @ sO:N] という歌詞の場合)

コンをマウスで操作することで指定することができる．また，図 -3 の下半分に示されるように，各種合成パラメータを直接事変的に操作することができるようになっている．

スコアエディタに入力された情報は，専用の MIDI メッセージに変換されて合成エンジンに送られる．その MIDI メッセージは，通常の Note On/Off は使用せず，前述したような母音開始位置によるタイミング合わせが可能な形式となっている．すなわち，合成に必要なすべての情報（Note On/Off に相当する情報さえも）を事前にディレイ情報付きで NRPN（Non Registered Parameter Number）のフォーマットで送っている．具体的には，「今から D[ms] 後に，ノート番号 n，duration が d[ms] で歌詞が L であるような音符を鳴らしなさい」という内容を合成エンジンに送っている．

(b) 歌手ライブラリ

歌手ライブラリは実際の歌手の歌唱データから取り出した音声素片を集めたものである．素片の単位としては，歌声としての性質上，多音素連鎖を用いると処理が複雑化するため，現時点では二音素連鎖と伸ばし音のみとなっている．対象となる言語で可能性のあるすべての C-V，V-C の組合せと母音，鼻音の伸ばし音が含まれる．

音声素片用の録音では，効率的に素片が収集できるように考案された専用の歌詞を歌手に歌ってもらう．声域によって声質が変化するので，収録は複数のピッチで行う．もちろん，収録するピッチの数が多ければ多いほど合成音のクオリティ向上が期待されるが，歌手への身体的，心理的な負担を考慮して，ある程度のところで妥協が必要となる．

収録されたデータは音素セグメンテーションおよび使用する領域のセグメンテーションを自動的に行い，人間の手によるチェックと修正を経て完成される．

(c) 合成エンジン

スコアエディタが出力する MIDI メッセージに含まれる音符，歌詞，表情その他の情報に従って，合成エンジンは必要な音声素片を歌声ライブラリから取り出し，連結して合成する．その際の素片の使用タイミングは，前述したような母音の位置と音符開始位置に合うような調整がなされる．すなわち，合成エンジン内部には図 -4 に示されるような，内部スコアがあり，C-V という素片の V の開始位置と音符開始のタイミングが合うような素片の位置調整が行われる．

合成スコアには，各時刻でのピッチや各種合成パラメータの変化も描かれ，合成時に参照される．ピッチに関しては，指定された音符とアタック，ビブラートのパラメータをもとに，ピッチのカーブが内部的に計算され，

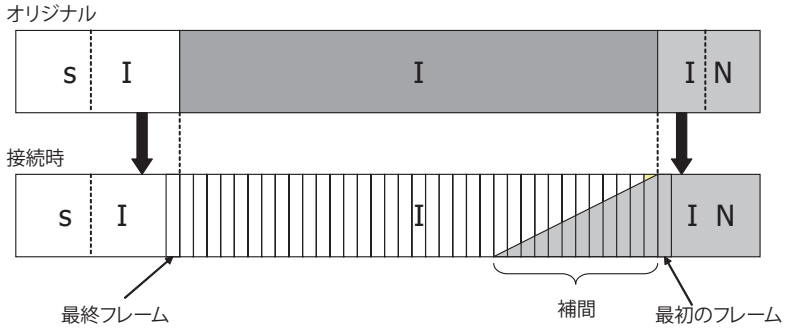

図 -5　スペクトル包絡の補間

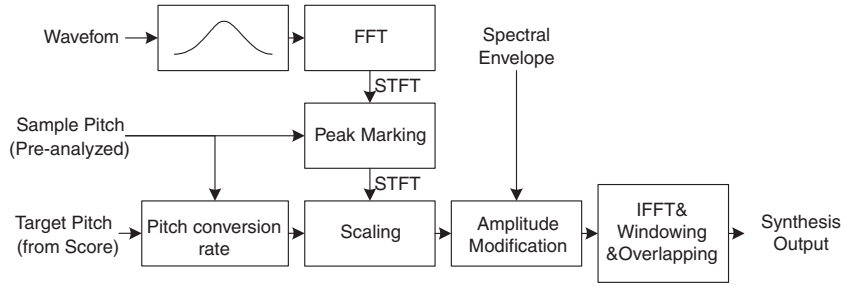

図 -6　合成エンジンの信号処理ブロックダイヤグラム

合成スコアに格納される．

素片の連結時には，素片のピッチを所望のピッチに変換する必要があるが，2 つの素片の接続部分のピッチを合わせたとしても，単純に接続するだけでは 2 つの素片の音色の差がノイズとなって現れる．素片連結時には音色も合わせ込む必要がある．

Vocaloid では，伸ばし音の区間で隣り合う二音素連鎖のスペクトル包絡を補間することで音色の合わせ込みを行っている．図 -5 にその例を示す．図 -4 では "sing"（[sIN]）という歌詞の伸ばし音のスペクトル包絡は，伸ばし音に先行する二音素連鎖すなわち [s-I] の最終フレームと，伸ばし音後の二音素連鎖 [I-N] の最初のフレームのスペクトル包絡を時間的に補間することで求められる（局所的なスペクトルは伸ばし音の音声素片のものを使用する）．これにより原理的に連結部分で音色の急激な変化が発生しないようになっている．二音素連鎖区間は，素片に含まれる各フレームのスペクトル包絡および局所的なスペクトルがそのまま使用される．

スペクトル包絡の補間をしやすくするため，かつ合成音の音色をユーザがある程度コントロールできるようにするため，実際にはスペクトル包絡を一定の中心周波数，バンド幅，強度を持つ 2 次のフィルタがいくつか加算された形で表現されている．

所望のピッチ，スペクトル包絡が決まったので，素片をそれに合わせるように変換する．変換は図 -6 のブロックダイヤグラムに示されるような処理により行われる．

ピッチ変換およびスペクトル包絡の調整は，図 -7 のようにスペクトルを周波数軸上でスケーリングし，ピーク部分の強度を調整することによって行われる．

ピッチ変換時には，倍音に相当するピークの近傍のスペクトルの形状はできるだけ元のものを保つように，非線形にスケーリングが行われる．このとき，倍音の周波数が完全に整数倍になっていると仮定し，i 番目の倍音に相当する周波数の位相に対して，以下の式で補償が行われる．

$$\Delta\varphi_i = 2\pi f_0 (i+1)(T-1)\Delta t \qquad (1)$$

ただし，T は f_{0T}（所望のピッチ）と f_0（素片のオリジナルのピッチ）の比であり，Δt はフレーム長である．

周波数領域でのピッチ変換と音色の調整の後，IFFT と Windowing & Overlapping を行うことで合成音声が得られる．

歌声合成の今後

歌声合成の品質は今後ますます向上していくだろう．

その中で，特に歌い方や表現のモデリングは自然な歌声を作り出す上で特に重要になってくるだろう．特にさまざまな音楽スタイルに合った歌い方や，特定の人の歌い方や癖を再現できるようなモデルが期待される．また，歌声合成システムユーザの表情付けのための作り込みの労力を減らすための手法として VocaListener というシス

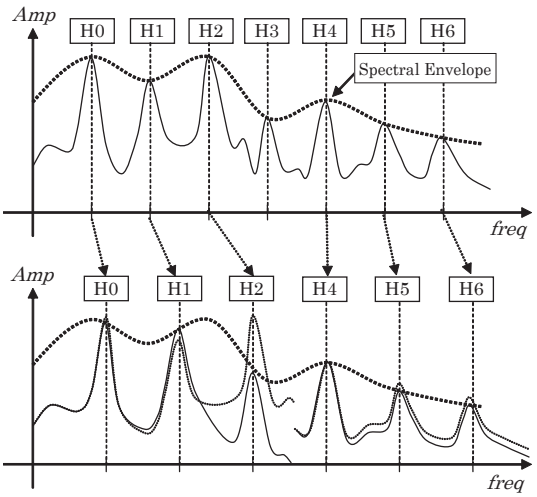

図 -7　ピッチ変換およびスペクトル包絡の調整

テムが提案されている[4]．これは人間の歌唱に合わせるように歌声合成システムのパラメータ（ピッチやダイナミクスなど）を自動的に調整するものである．簡単に高品質の歌唱を得られるツールとして早い時期の実用化が期待される．

合成方式で言えば，テキスト音声合成で注目されている HMM 合成方式を歌声合成にも応用する試みが行われ，一定の成果が出ている．この手法は，元の歌声の提供者の声質や歌い方を再現することができる．また話者適応により簡単に別の歌声の提供者の声に変換することも可能なので，今後も注目される技術である[5]．

本稿では，歌声合成とは，楽譜情報（歌詞と音符）またはそれに相当するものを入力とし，歌声を出力するものとして扱ってきたが，それ以外にも話し声を歌声に変換するようなシステムも提案されており，高品質な歌唱が得られる手法として注目されている[6]．

技術的な発展とともに，歌声合成技術のアプリケーション開拓もますます進んでいくこととなろう．Vocaloid 合成エンジンをサーバ上で動作させ，SaaS（Software as a Service）として提供する NetVocaloid と呼ばれるサービスも実際に運用されており，携帯電話向けにサービスが開始されている．携帯電話以外への応用も期待されている．

アプリケーションを考えるとき，合成音に当たり障りのない平凡な声ではなく，特徴のある声であることが要求される場合が多い．ある特定の人の声であることが求められる場合もある．しかしその場合，合成音は声の提供者の声とは，どんなに近づけたとしても結果として似て非なるものであるため，利用者側には逆にギャップを感じてしまい，アプリケーションとしては失敗しやすい．これは元の声の提供者の声の代用として歌声合成システムを使おうとしていることから来る悲劇である．元の声の提供者と合成音をある程度切り離すような工夫が求められる．

合成音の品質向上とともに，商業音楽シーンで歌声合成が利用される機会もますます増えていくことであろう．現状では，「ニコニコ動画」などの動画投稿サイトで人気になったコンテンツがメジャーレーベルによって CD 化されるという流れであるが，はじめから歌声合成を使用することを意図して制作されるコンテンツも増えていくことであろう．

今後歌声合成が一般化し，品質も向上していくときに注意しなければならないのは，声の提供者の権利である．声の提供者には一定の著作物（すなわち収録するための楽譜，歌詞）を歌っていただくということから，現状では声の提供者は著作権上の隣接権（実演家の権利）の保持者であるとみなし，隣接権の中の財産権の部分は金銭を対価として譲渡（場合によってはロイヤリティの支払い），人格権に関する部分は行使しないことに同意していただくという，契約ベースの処理が合理的だと考えられるが，歌声合成のクオリティが今後ますます向上するにつれて，合成音は誰のものなのか，声の提供者の権利はどこまで及ぶのか，また合成音声の場合，楽曲の「歌手」とは誰になるのか，ということについてコンセンサスが必要になってくるだろう．

参考文献

1）Kelly, J. et al. : Speech Synthesis, Proceedings of the Fourth International Congress on Acoustics, pp.1-4 (1962).
2）McAulay, R. et al. : Speech Analysis/Synthesis Based on a Sinusoidal Representation, IEEE Transactions on Acoustics, Speech and Signal Processing 24(4), pp.744-754 (1986).
3）Rodet, X. : Time-Domain Formant-Wave-Function Synthesis, Computer Music Journal 8(3), pp.9-14 (1984).
4）中野 , 後藤 : VocaListener : ユーザ歌唱を真似る歌声合成パラメータを自動推定するシステムの提案 , 情報処理学会 研究報告 2008-MUS-75 Vol.2008, No.50, pp.49-56.
5）酒向 他 : 隠れマルコフモデルに基づいた歌声合成システム , 情報処理学会論文誌 , Vol.45, No.3, pp.719-727 (Mar. 2004).
6）Saitou et al. : Vocal Conversion from Speaking Voice to Singing Voice Using STRAIGHT, Proceedings of Interspeech 2007, pp.4005-4006.

（平成 21 年 6 月 30 日受付）

剣持 秀紀　kenmochi@beat.yamaha.co.jp

1967 年生まれ．1993 年京都大学大学院工学研究科電気工学第二専攻修士課程修了．同年ヤマハ（株）入社．1996 年エル・アンド・エイチ・ジャパン（株）出向．1999 年ヤマハ（株）に復職．以降歌声を含む音声信号処理に関する研究開発に従事．