高品質音声分析変換合成システム STRAIGHT を用いたスキャット生成研究の提案

河原英紀[†,††,†††] 片寄晴弘[†]

音楽としての歌唱の魅力は，歌詞をともなうことに多くを負っているといわれる．しかし，歌詞の理解できない外国語の歌唱であっても，楽器としての人間の声の魅力を楽しむことができることも事実である．ここでは，楽器としての声そのものの魅力を楽しむスキャット，ヴォーカリーズ，口三味線，鼻歌等を対象として取り上げ，音声処理技術を用いて，その魅力の分析，再合成，加工を行うシステムの開発を狙う一連の研究構想を提案し，実現技術の予備検討結果を紹介する．具体的には著者らが開発している高品質音声分析変換合成システム STRAIGHT をエンジンとして利用し，基本的な反射弓を修飾する発声制御モジュール，韻律制御モジュール，音楽情報処理モジュール，インタラクション制御モジュール等を逐次更新していく生態学的枠組みに基づく開発戦略を提案する．様々な研究者が，このようなシステムの実現を意識して研究を進めることは，計算機音楽の範囲を拡大するだけではなく，音声に含まれる非言語情報やパラ言語情報の処理技術に対する有力なベンチマークの機会を提供するものと考えられる．

Scat Generation Research Program Based on STRAIGHT, a High-quality Speech Analysis, Modification and Synthesis System

Hideki Kawahara[†,††,†††] and Haruhiro Katayose[†]

A research program to develop a versatile system for analysis, manipulation and generation of a specific vocal music genre; scat, vocalease, *kuchi-jamisen* and humming, is introduced. One of the major aim of the program is to explore why vocal music is still attractive, even if their lyrics are not intelligible when they are sung in a foreign language. This may sound peripheral to the usual belief that lyrics is the central charm point of vocal music. However, we argue that this type of research is indispensable for understanding roles of non-linguistic and para-linguistic components in speech and vocal music. The proposed program uses STRAIGHT as its central analysis, modification and synthesis engine, and will refine its constituent modules like voicing control, prosodic control, musical information processing, interaction control, and so on, organized as modifiers of the basic reflex arc, in an evolutional and developmental process. This research program, that can be understood as a global load-map for various individual research projects, provides a unique common ground for benchmarking non-linguistic and para-linguistic processing algorithms as well as a wide variety of opportunities in computer music applications.

1. はじめに

すでにリタイアした歌手の歌声を聞きながら,「この人にあの歌を歌ってもらえたら...」と想うことがある．華やかな技巧をちりばめた曲を聞きながら「この曲を自分が歌えたら...」そう想うこともある．それらの願いは，多くの場合，時期的な制約，能力的な制約，時間的な制約のためにかなえられることはない．しかし，この状況は変わりつつある．計算能力と記憶容量の爆発的な増加を背景とすれば，聴覚情報処理，音声情報処理（特に非言語情報およびパラ言語情報），音楽情報処理，インタラクション技術，運動制御の計算理論等を総合することで，これらの願望を満たすシステムを実現することは工学技術の射程に入りつつある．そのような個人的な願望の充足のために高度の技術資源を集中することは，効率と速度の 20 世紀的価値観からは非常識なことであろう．しかし，人間の世紀となるべき 21 世紀の入口に立った現在，価値観の

† 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University
†† ATR 人間情報科学研究所
Human Information Science Laboratory, ATR
††† CREST
Core Research for Evolutionary Science and Technology

中心を個人の幸福に置く技術体系の構築を真剣に考えてもよいのではないだろうか．

ここでは，そのようなシステムを実現するための1つの里程標として，スキャットを生成するシステムの開発プロジェクトを提案する．提案するプロジェクトの中核には，筆者らによって開発が続けられている高品質音声分析変換合成システムSTRAIGHT[1)~3)]が据えられている．本論文では，プロジェクト提案の背景を明らかにするとともに，STRAIGHTを応用して，いくつかの要素技術に関する予備検討を行った結果についても報告する．

2. STRAIGHTの概要

STRAIGHT[1),2)]は，音声分析合成技術の原点である1939年のVocoder[4)]の現代版である．Vocoderと同様に，入力された音声は，基本周波数等の音源情報と，調音器官により形成される声道等の特性を表す滑らかな時間周波数特性に分解される．滑らかな時間周波数特性は，音声合成時のスペクトル包絡を与えるために最小位相のインパルス応答として利用される．STRAIGHTの特徴は，基本周波数により与えられる信号の周期性を積極的に利用した適応的な分析を行っているところにある．そのため，通常の分析で問題となる時間周波数特性への基本周波数の干渉は，ほぼ完全に取り除かれている．また，基本周波数情報自体も，wavelet分析の瞬時周波数に現れる不動点を利用した新しい方法により，安定に高精度に抽出される[5)]．これらのパラメータは，実数の組として表されており，それぞれ独立に操作することができる．このように，高精度で相互の干渉のないパラメータの組として音声を表現することにより，再合成音声の品質を損なわずに，広範にパラメータを変換することが可能となる．なお，STRAIGHTの出発点であった周期性は，現実の音では不完全に成立しているにすぎない．周期信号から非周期信号までを連続的に表現し制御するために，音源の群遅延特性を操作していることもSTRAIGHTの特徴の1つである[6)]．

人間の発声機構を精密に解析し模擬することによって高品質の合成音声を作成しようとする研究がある[7)]．STRAIGHTは，そのような研究とは逆の方向から，人間が聴覚でとらえる音の属性に沿って音を分析・操作・再合成することを狙っている[8)]．

3. 目標の設定

ここで実現を目指すシステムでは，特定の言語に依存する歌詞を扱わないこととする．歌詞を扱わないことにより，声と個人性，表現，情緒，感動との関係を，より直接的に追求することを狙う．システムのアーキテクチャの選択にあたっては，計算機メタファに基づく安易な工学的モジュールへの機能分割は行わない．むしろ，人間のアーキテクチャに倣い，低次の反射弓とそれを修飾する高次の反射弓や計画–制御モジュールが層状に積み重なった並列階層システム[9)] [pp.396–400]として構成することを狙う．いわば，進化と発達によって形成された脳のアーキテクチャに倣うのである．以下では，相互作用を通じて発達するシステムとして発声・発話を概観する[10)]．

発声は，母親等の養育者との相互作用から始まる．言葉を発するに至る前に，マザリーズと喃語での相互作用が長く続く．ここでは，養育者の声と自分の声の同一性の知覚が相互作用と発達を導く．この期間を通じて，基本周波数の変化パターンと母音のような音によるバリエーションを中心とする喃語は，徐々に複雑さを増す調音運動をレパートリに加え，言葉に至る．こうして最終的に獲得されるレパートリは，それぞれの言語環境に固有のものとなる．しかし，言語依存部分の下には膨大な共通の基盤がある．この基盤は，発声と調音を制御する基礎的な機構から構成されており，言語に依存しない生物学的，生態学的拘束の下に形成される．スキャットの生成は主にこの基盤を利用し，子音や母音等のレパートリは，言語の語彙的，統語的，意味的内容からは切り離されて音の素材として利用される．

このようにして生成されるスキャットと同じものが生成できるようなシステムを工学的手段で実現することを目標とする．開発戦略としては，組織的ダウングレードとでも呼ぶべき方法と，発達過程の模倣という2つの戦略を採用する．

3.1 高い品質の維持

人間の知覚のように高度に非線形なシステムを調べる場合，目標とする刺激から大きく外れた試験用の刺激を用いることは，誤った結果を導く．たとえば，音色や表情の細かなニュアンスにも重要な要素が含まれているような歌唱を，劣悪な品質の合成音声で再現して評価することは不適切である．人間による評価が必要な場面では，研究の各段階でつねに高い品質を保つことが必要なのである．組織的ダウングレードは，そのような要請を満たす研究戦略として考えられたものである．

まず，十分に鑑賞に耐えるスキャットを，大量の録音素材からSTRAIGHTを用いた分析変換合成により作成するところから出発する．具体的には，様々な条件（基準となるピッチ，速度，音程，音節，唱法，表情

の各々について，生成の対象となるスキャット中に出現するものをカバーできる水準を用意する）で発声された旋律の断片を素材として，目標とするスキャットの構成要素に最も近い断片を STRAIGHT を用いて変換し滑らかに接続することから出発するのである．これは，コーパスに基づいた音声合成法を STRAIGHT を用いて拡張し，歌唱に応用することに相当する．次いで，高い品質を保ったまま，録音素材に含まれる水準を減少させていくことによりサイズを段階的に縮小させて，STRAIGHT の合成器に渡すパラメータの時系列を，モデルを用いた補間と外挿によるものと置き換えていく．初期の段階では，STRAIGHT を用いて鑑賞に耐えうるスキャットを生成するためには，変換と接続およびモデルから得られるパラメータの時系列に，さらに手作業で多くの修飾を加える必要がある．モデルの詳細化と洗練のステップでは，前述の補間と外挿に加え，これらの修飾の内容をルールとして抽出してサブシステムのプログラム化された動作で置き換えることで☆，さらに用いる録音素材を削減する．

このステップを繰り返すことにより，最終的には，少量の録音素材および楽譜と演奏意図を与えるだけで，鑑賞に耐えるスキャットを生成するシステムを実現する．このようにして作成されるシステムの構造にはかなりの自由度があるが，ここでは，人間のアーキテクチャに倣い，歌唱に関わる器官のダイナミックスと逆ダイナミックスモデルおよび聴覚の応答特性を要素として含む形に構成することを狙う．楽譜と演奏意図情報は，目標とする歌唱を実現するために必要な発声発話器官の制御指令を表す目標軌跡と，それら器官のダイナミックスを制御するパラメータの系列からなる内部情報に変換されて，演奏に利用される．

発達過程の模倣は，組織的ダウングレードによって作成されたシステムから出発する．ここでは，内部で利用する適切なダイナミックスを制御するパラメータの系列と目標軌跡を，強化学習の枠組みで，教師の演奏を模倣し練習することで形成するシステムの実現を目指す．

以上をまとめて実現すべきシステムの概要を**図 1** に示す．呼吸に関する最も低次の反射弓が省かれており，順システム逆システムから構成される細部も省略されているが，人間のアーキテクチャ[9] [pp.396–400] に倣ったシステムの実現を狙ったものとなっている．

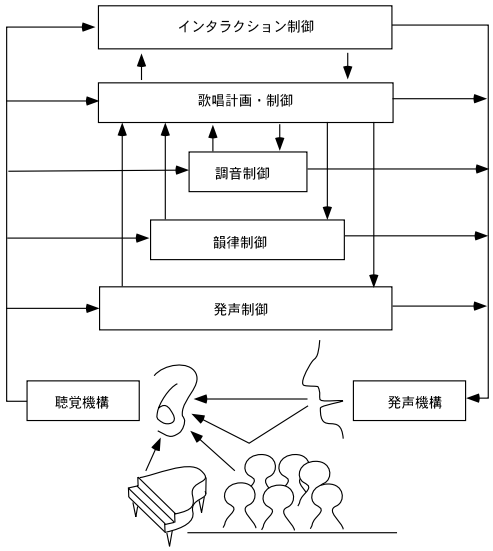

図 1 スキャット生成システムの概念モデル
Fig. 1 Conceptual outline of the target scat generation system.

4. 技 術 課 題

前の章で示した目標を実現するためには，様々な技術的課題を解決することが必要となる．以下では，それらの課題を枚挙するとともに，現状について概観する．残されている技術的課題の解決と STRAIGHT の構成要素との関わりならびに解決策の実装例については，章を改めて論ずる．

4.1 基本周波数の精密な抽出と制御

まず，最初は，システムを構築するための要素に関する課題である．単純なパルス音源を用いる Vocoder 型の音声合成システムでは，高い基本周波数の音声のピッチ☆☆を音楽に必要な精度で制御できないという問題があった．これは，標本化周期が基本周期に対して無視できない大きさとなることによる．正弦波合成型のシステム[11]では，問題とならない．パルス音源の場合であっても，標本化時刻と本来パルスの存在すべき時刻との差を補償するような直線位相特性の付与によって，この問題を回避することができる．

同様に，素材の基本周波数情報の抽出においても，自己相関に基づく方法では，標本化周波数が可能な分解能に制約を与える．また，基本周波数が時間ととも

☆ STRAIGHT の制御は，基本周波数，スペクトル包絡，帯域ごとの非周期性，群遅延等のパラメータの調整を通じて行われる．これらのパラメータは相互依存性が少ないため，修飾の結果は，どのような順序で操作を加えたかに依存しない．このことは，STRAIGHT 自体が手作業の内容の解析とルール化を容易にする（広義の）プロトコル解析に適した環境であることを意味する．

☆☆ ここでは，ピッチは知覚される心理量，基本周波数は信号の物理的属性を表すものとして，使い分ける．

に急速に変化する場合には，階段状の基本周波数の軌跡が得られるという問題がある[12]．

4.2 基本周波数制御の動特性の抽出と実装

概要で説明した開発戦略をとる場合，次に問題となるのは，基本周波数や調音を制御するモジュールの特性の解明である．ここでは，それらのモジュールを人間の機構と機能的に同型とすることを狙う．

歌唱音声の基本周波数は，楽器と比較するとはるかに複雑な軌跡を示す．しかし，少なくとも，伝統的な西洋音楽の歌唱では，離散的な表現である楽譜の旋律をそのまま楽器で演奏したものと同様な離散的な要素の組合せとして聴き取ることのできるような演奏も，それほど不自然ではなく可能である．ここでは，まず，離散的な『目標』が与えられた場合の基本周波数の軌跡を，物理的には離散的なものとさせない要因とその影響を概観する．

4.2.1 基本周波数制御の生理的制約

発声には多くの器官が関わっている[13]．基本周波数は，最終的には，声帯の振動速度により決まる1つのパラメータで表される量である．しかし，その制御は，喉頭周辺だけでも15種類の筋肉に依存しており[13] [pp.11–15]，呼気の供給に関与するものを加えると，20種類を超える過剰な自由度を持つシステムにより行われている．

このようなシステムにおいて，基本周波数を精密に制御することは困難な課題である[13] [pp.279–306][14),15)]．しかも，基本周波数を一定にさせない多くの要因が存在する．基本周波数を一定に保つには，呼気の排出にともなう声門下圧の低下，心拍による声門下圧の変動や脈波による声帯質量の変動[16]，筋肉への指令パルスの確率的変動等の影響を補償することが必要となる．

4.2.2 声道の狭窄の影響

子音の調音では，通常，声道の途中に狭窄あるいは閉鎖が形成され呼気流が妨げられる．呼気流への抵抗は，等価な声門下圧の低下につながる．その結果，多くの場合，基本周波数は子音の部分において低下する．これらの言語情報の干渉を受けて，基本周波数軌跡には，計画された軌道からの局所的な逸脱が重畳することになる．

4.2.3 聴覚フィードバックの影響

発声時の基本周波数は，聴覚によりつねにモニタされて修正されており，その動特性が測定されている[17)~19)]．モニタされた基本周波数の情報は，速度が遅いが大きなループゲインを有する系と，速度は早いが小さなループゲインを有する系とにより，並列にフィードバックされ，目標からのずれを補償しているようである．また，早い系であっても，基本周波数の変化に反応するまでには，全体として100 ms以上のむだ時間を含んでいる[17]．このような系が，フィードバックのみによって制御されていると考えることは困難である．発声においては制御が逐次的なのではなく，逐次的なフィードバック誤差学習によって制御対象の動特性の把握と調整が行われていると考えた方がよい．具体的には，大脳による調整への関与を得て，順モデルと逆モデルが形成されて，制御そのものは前向き制御で行われると見るのである．なお，最近の研究は，目標値そのものも，聴覚からのフィードバック情報に基づいてゆっくりと修正されていくことを明らかにしている[19]．

4.2.4 ピッチ感覚の時間特性

上で説明した聴覚によるモニタに関して，未解決の問題がある．ピッチ感覚の時間特性である．歌唱の制御に対して運動制御の計算理論の適用が困難であるのは，目標と実現との比較が行われるレベルと表現が明らかではないことによる．物理量である基本周波数に対応する心理量であるピッチの知覚は，数100 msという大きな時定数を持つ遅いプロセスである（たとえば文献20) [pp.53–57]）．しかも，基本周波数の情報が利用可能になるまでには，基本周期が5回繰り返す程度の処理時間が必要であり[17]，逐次制御には間に合わない．基本周波数が早い速度で変化する場合のピッチ知覚（特に歌唱に関連するような）については，比較的単純な軌跡についての組織的な検討[21),22)]が開始されたばかりである．なお，基本周波数の周波数変動の中の十数Hz以上の速度で変化する成分は，ピッチ情報としてではなく，声に自然性を与える成分としての役割を担っていることが示唆されている[23]．

4.3 基本周波数とスペクトル包絡の相関

基本周波数が変化すると，音源波形と声道形状の双方に依存するスペクトル包絡は，いくつかの要因により変化する．それらは，基本周波数の調節メカニズムに起因する構造的なものであったり[24]，局所的な声門の開閉比率の変化にともなう音源スペクトル形状の変化であったり，聴覚フィードバックに基づく響きの局所的な調整によるものであったりする．自然な歌唱音声の生成には，これらを規則として抽出し，補間や外挿に取り入れる必要があろう[25]．

4.3.1 包絡ピークと調波の相関

歌唱では，できるだけ音源から供給されるエネルギーが効率良く音となって放射されるよう，スペクトル包絡のピークは，基本周波数の調波の位置に調整される傾向がある[13] [pp.231–232]．

4.3.2 基本周波数とスペクトル傾斜の相関

同 1 人物であっても異なる周波数領域では，異なった発声法を用いる．それらの発声法は，異なった声帯振動様式に結び付いており，音声のエネルギを供給する声門での駆動条件の変化として実現されている．これらの差違は，音源スペクトルの大局的傾斜に大きな影響を与える．

4.3.3 歌唱フォルマント

オペラ歌手の歌唱の分析で発見された 2,000 Hz〜3,000 Hz 付近でのエネルギーの強調は，歌唱フォルマント（singer's formant）と呼ばれている[13] [pp.239–241]．クラシックの楽曲の歌唱においては，このようなスペクトル包絡の変型を模擬することも必要となる．

4.3.4 基本周波数の変換と個人性の保存

ここまでに示したような基本周波数とスペクトル包絡との様々な相関は，スペクトル包絡を変形し，同 1 人物であっても異なった基本周波数/歌唱法に属するスペクトル包絡の同一性/相似性は崩れる．しかし，そのような変形にもかかわらず，多くの場合に聴取者は同 1 人物の発声に一貫性を感ずることができる．話し声に関して蓄積されてきた，スペクトル包絡等の声道特性および基本周波数等の音源特性と話者性との関係[26),27)]が，歌唱の場合のような大きな変形の場合にどのように成立しているかを明らかにすることは今後の課題である．

4.4 表情とスペクトル包絡および音源の相関

もう 1 つ必要な操作軸は，基本周波数，個人性を保ったままの，表情の変化である．声門の閉止部分の長さと開放部分の長さの比と音色との関連についての検討は存在する[13)]．しかし，それ以上微妙な『明るさ』や『柔らかさ』『透明感』『だみ声』等の属性．さらには感情との関連の検討が必要である．

4.5 調音の選択

速度の大きな旋律をスキャットで歌う場合,「ダバダバ〜」「ドゥブドゥブ〜」のように子音と組み合わせ，さらに 2 種類の異なった子音を交互に挟むことがよく行われる．ゆっくりとした旋律の場合には，同一母音の繰返しや「ラ〜ラ〜」のように子音と組み合わせる場合でも同一のものを繰り返す傾向がある．この傾向の背景にある規則を抽出し，旋律の局所的速度に応じて，適切な子音の組を自動的に選択する仕組みを組み込む必要がある．

4.6 演奏意図のモデル化

意識の脳神経基盤は，いまだに決着の着いていない困難な問題である．ここでは，意識は，自己の（観察可能な）内部状態を観測した結果として作り上げられる『説明』であるとする立場をとる．意識は，脳内で実際に進行している膨大な情報処理には直接関与することができず，結果として表出された（広義の）行動のみに基づいて出現するものであるとするのである．

演奏意図☆は，この文脈の下では，演奏者の意図を観客が受け取るときに，目的とした意図に受け取られるように演奏するための修飾の方法に付与されたラベルである．聴衆と演奏者が共通の「意図から演奏（音として出現した）への順モデル」を持つ場合には，演奏者の意図があたかも直接的に観客に受け取られることが可能となる．歌唱の場合には，器楽演奏[29)]と比較すると，生物学的に拘束される部分が多いため，演奏者の意図と観客が受け取るものには共通部分が大きくなることが期待される．

しかし，意図のモデル化には，方法論的な問題がある．前節で説明した「ラベル」は直接的には観測できない仮説的な実体である．アンケートを用いることによって，間接的に，事後に，言語的解釈を経た後の結果として報告を得ることは可能である．演奏者や観客が，演奏中に意図し受け取ったラベルをアンケートへの記入時まで保持でき，かつその（本来は連続的な）ラベルと（本来離散的である）言語的表現とが 1 対 1 に対応づけられるのであれば，この方法でも，意図のモデル化のための基礎データを得ることは可能である．しかし，一般的には，このような条件が成立していることを期待することはできない．むしろ，作成したシステムに，様々な意図で演奏した資料を模倣するように学習させた場合の制御パラメータの既定値からの修飾量のクラスタ分析と，それらクラスタと演奏者/観客の生理的指標やアンケート結果との相関の解析を通じて，明らかにすべき問題であろう．

4.7 インタラクションのモデル化

自分自身，伴奏，観衆等の状態を把握しインタラクションを制御するモジュールは，歌唱システムとして完結するためには，不可欠の要素である．しかし，この問題は，一般的な音楽演奏におけるインタラクションの枠組みで議論すべきであり[28)] [pp.206–217]，ここでは論じない．

5. STRAIGHT における技術的課題の解決

技術的課題で枚挙した問題点のいくつかは，現在の STRAIGHT の実装において解決されている．その他

☆ この言葉の示す概念は多数のレイヤを有しており，この言葉は様々なレベルで使われる[28)] [pp.199–201]．ここでは，表現様式や手段ではなく，それらの操作を生み出す原因となる高いレベルを指す言葉として用いている．

のものについては，簡単な拡張により解決できる課題から，綿密な研究を必要とするものまで，広範にわたっている．この章では，STRAIGHTに直接関わる解決策のいくつかについて概要を説明する．前の章であげた広範な技術課題の全体は到底1つの研究機関だけで解決できるものではない．今後，問題意識を共有した多数の研究者の協力により1つ1つ綿密な検討が進められていくべきものである．

5.1 基本周波数の精密な抽出と制御

Vocoder型の音声合成システムでありながら，STRAIGHTにおける基本周波数の制御は，標本化周波数による分解能の制約を受けない．これは，群遅延操作を行う音源生成機構を用いているために，パルス位置の制御にその機構を流用できるためである．基本周波数の抽出における分解能の問題は，低域フィルタと自己相関のピーク近傍での放物線近似によって回避することもできる．しかし，本質的な解決は，STRAIGHTにおける実装[2]やその後の改良[5]で用いている瞬時周波数に基づく方法による必要がある．階段状の基本周波数の軌跡の問題についても，それらの方法では一般化されたピッチ同期分析によって本質的に解決されている．

5.2 基本周波数制御の動特性の抽出と実装

著者の1人によって発明された変換聴覚フィードバック[17),30)]を用いることにより，聴覚フィードバックの影響を含めたシステムとしての基本周波数制御の動特性を測定することができる．このシステムは，2 Hz以下の領域で働くフィードバック制御系と，変動の補償方向に働く150 ms程度の時間遅れを含む系の並列接続として近似できるような応答を示す．このような性質を有する基本周波数の制御系の実装にはいくつかの代替案が存在する．

5.2.1 基本周波数軌跡の計画

言語音声の基本周波数軌跡の計画に関しては，Fujisakiによる先駆的な試みがあり[14)]，そのモデルは現在も広く用いられている．ただし，これは習熟した言語行動に関するモデルであり，素人の段階から玄人に至る様々な習熟段階の歌唱にまで適用してよいか否かは明らかではない．また，(may not be exactly critically-damped)[14)] [p.234]との保留をつけられながらも文章音声のよいモデルであるとされている臨界制動二次系は，少なくとも歌唱における基本周波数の動特性の記述のためには拘束が強すぎるおそれがある☆．ここは，既存のモデルを流用するのではなく，自分自身を含んだ環境と相互作用するシステムとして歌唱をとらえ，基本的なレベルから議論を再構築すべきであると考える．急速に発展している運動制御の計算理論の枠組み[9),33),34)]をこのレベルに適用することは，本質的な理解と適切なモデルの構築のための鍵となろう．変換聴覚フィードバックによる測定結果は，この枠組みと整合する情報を提供するものである．

5.3 音源の非周期成分の制御

高品質な合成音声の作成においては，音源に含まれる非周期成分の抽出と制御が必要である．音源の群遅延パラメータの制御は，STRAIGHTにより初めて音声合成に導入され[6)]，自然性の向上に大きく貢献した．予備的な検討によれば，歌声の場合には，話声と比較するとはるかに小さな群遅延の広がりが適していることが観測されている．群遅延のパラメータは，音源の帯域ごと，基本周期内位相ごとの非周期性を制御するパラメータ群の1つである．これらのパラメータは，STRAIGHTのために開発された2つの音源情報抽出手法により求められる．1つは，周波数領域での写像の不動点に基づく方法であり[5),35)]，もう1つは，時間領域での写像の不動点に基づく方法である[36)~38)]．様々なクラシック唱法について，これらの分析方法を応用し，STRAIGHTを用いて合成する方法についての議論は文献39)を参照されたい．

5.3.1 特異な声帯振動への対応

いわゆるクラシック音楽における歌唱では，周期性がはっきりとし，(ビブラートを除けば）安定した基本周波数を持つ声が用いられる．しかし，日本の演歌や様々な伝統歌唱，ポピュラー音楽の領域では，だみ声や叫び声，気息性の声等，多様な表現が用いられる．周波数領域の不動点による方法[5)]や，時間領域の不動点による方法[37),38)]を用いると，それらの多様な表現に特徴的なパターン☆☆を認めることができる．

6. 実音声の分析と加工例

以下では，上で紹介した解決策の中から基本周波数制御の動特性を変換聴覚フィードバックによる測定結果を利用してモデル化する方法を取り上げる．ここでは，実際の歌唱音声の基本周波数軌跡の分析結果を例にとり，4章で指摘した問題が具体的にどのように表現されているか，モデルを用いることで，素材となる音声がどのようにSTRAIGHTで加工されて目標と

☆ 先行研究[31),32)]とは，この点で異なった見方をしている．

☆☆ 後で示す『だみ声』の例（図6）では，低いC/Nの領域を持つ不動点の軌跡が1オクターブの間隔で並行して走っている．通常の母音の場合には，安定して低いC/Nを示す領域をともなう不動点の軌跡は1本だけとなる．

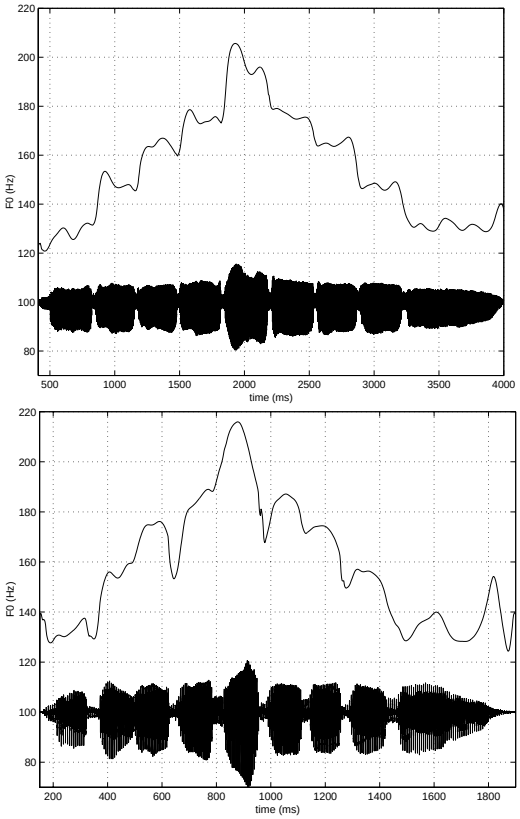

図 2 男性の歌唱音声の F0 の例
（上段：遅い演奏，下段：速い演奏）
Fig. 2 F0 trajectory examples by a male singer.
(upper: slow performance, lower: fast performance)

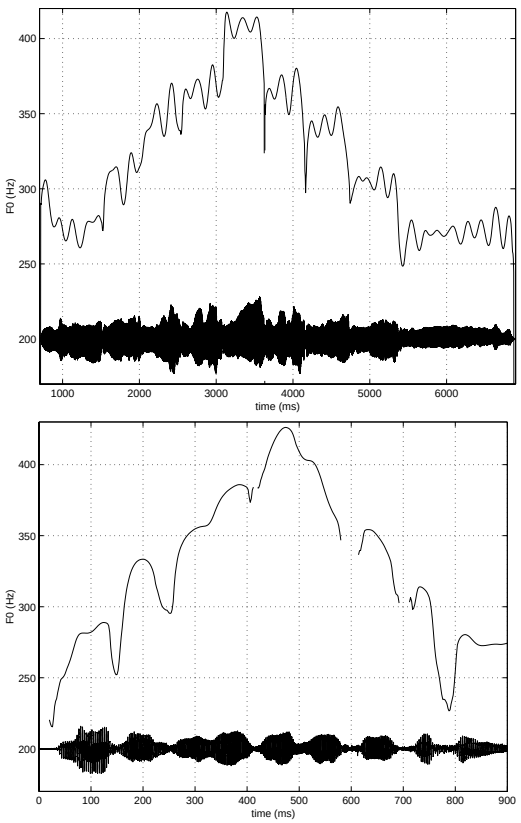

図 3 女性の歌唱音声の F0 の例
（上段：遅い演奏，下段：速い演奏）
Fig. 3 F0 trajectory examples by a female singer.
(upper: slow performance, lower: fast performance)

するスキャットが合成されるかについての例を示す．

図 2，**図 3** は，異なった速度で歌われたスケールの基本周波数の軌跡である☆．10 年以上の西洋音楽の歌唱の経験のあるアマチュアの男女による歌唱である．上段は，ゆっくりとした試行の例を示す．下段の例の採取では，階段状のスケール感を保持できる範囲で，できるだけ速く歌うように教示した．スキャットに用いた言葉は，ゆっくりとした試行の場合には「ララララ〜」，速い試行の場合には「ラバラバ〜」である．なお，女性のゆっくりとした歌唱には強めのビブラートがかかっている．

データの収録には，Sony ECM-23F マイクロフォンを用い，44,100 Hz，16 bit の標本化を行った．基本周波数の抽出は STRAIGHT に実装されている周波数領域での不動点に基づく方法を用いた．ゆっくりとした試行の場合には，5 ms，速い試行の場合には，2 ms ごとに基本周波数を求めた．求められた基本周波数の誤差は，母音中央部では 0.5 Hz 以下である．ただし，/b/の子音部では S/N が低下し周期性も崩れるため，数 Hz 以上の誤差が発生している．また，速い試行の場合には，子音部における声道狭窄にともなう基本周波数の局所的な下降が顕著に認められる．

これらの図より，基本周波数を制御するシステムは，臨界制動二次系よりもはるかに制動不足の状態にあることが分かる．なお，これらの試行は，速い試行の場合でも，正しい音程を等間隔で歌っているように聞こえる☆☆．速い試行の基本周波数の軌跡からは，これらがどのような機構で，等間隔に正しい音程で演奏されたスケールであるように知覚されるのかを説明することが困難な課題であることも分かる．この問題は，ピッチ知覚機構の研究と並行して進める必要のある課

☆ これらの図示，モデルの計算は，すべて対数周波数を用いて行うべきである[14]．しかし，ここでは，便宜的に直線周波数を用いている．

☆☆ 幼児期から馴染んでいる西洋音楽のスキーマに基づいて，知覚レベルでのあいまいなピッチがスケールに乗るように解釈し直されているという可能性は排除できない．これらの試料は後述のウェブページに載せておくので，各自で判断いただきたい．

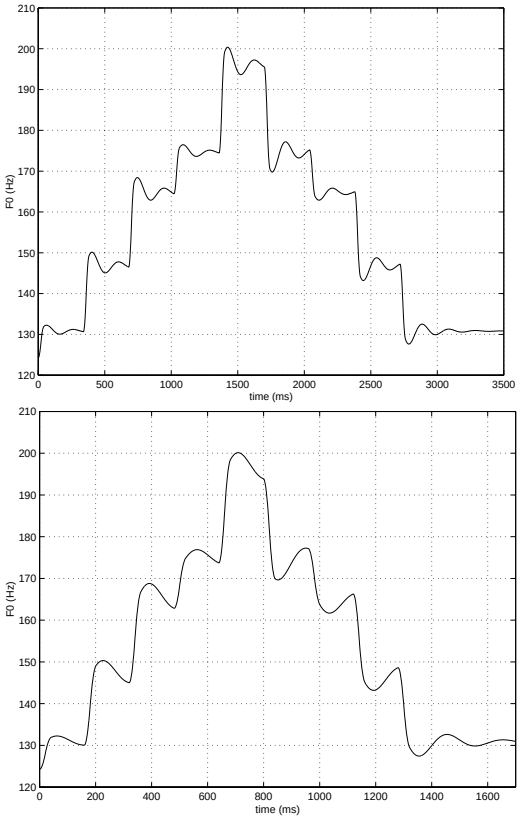

図4 モデルにより生成されたF0軌跡
（内部モデルによる前向き制御の場合）

Fig. 4 F0 trajectories generated by a preliminary model (with a feed forward control based on an internal model).

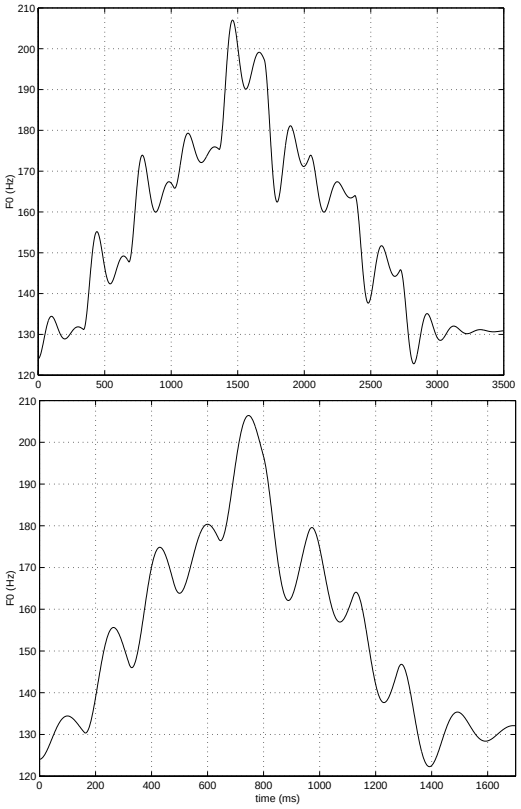

図5 モデルにより生成されたF0軌跡
（前向き制御がない場合）

Fig. 5 F0 trajectories generated by a preliminary model (without a feed forward process).

題である．なお，同一のスペクトル包絡，音源の非周期成分を有していても，基本周波数が階段状に音階のスケール上を移動するように変更するだけで，声らしさが失せて楽器のような響きとなることを特記しておく．

6.1 軌跡生成の予備検討

組織的ダウングレードにより録音素材の利用を削減させていった極限は，一定の高さで持続的に発声された素材からのスキャットの生成である．そのためには，4章で説明した要因のすべてについてのモデルを用意しなければならない．ここでは，すでにある程度の手がかりが得られている基本周波数軌跡の生成を例にとり，持続発声された母音からモデルとSTRAIGHTを用いて実音声の例で示したようなスケールの歌唱を生成する方法について説明する．

図4に，変換聴覚フィードバック実験[17),30)]により求められた音声の生成知覚における基本周波数制御モデルに，そのモデルを順モデルとして内部に持つ前向き制御システムの入力として階段状のスケールを入力して生成した基本周波数軌跡の例を示す．上段は，1つの音符の長さが340 msの比較的ゆっくりとしたスケールの場合，下段は，1つの音符の長さが160 msの速いスケールの場合である．ここでは，子音の狭窄による局所的な基本周波数の低下はモデル化していない．この予備的モデルでは，ピッチ知覚の周波数特性，運動制御における制御周期，音声の生成知覚システムにおける時間遅れを考慮している．しかし，内部基準の順応，筋紡錘からの内部フィードバックによる10 Hz付近の極，随意系を介したピッチの補償システムはモデル化していない．参考のため，前向き制御系を外した場合の基本周波数の軌跡を**図5**に示す．これは，いわば歌唱の経験がまったくない場合をシミュレートしていることに相当する．

前向き制御を取り入れたモデルによる基本周波数の軌跡は，実際の音声で観測された軌跡の特徴をとらえている．持続発声された母音をSTRAIGHTを用いて分析し，分析結果の基本周波数軌跡をモデルにより

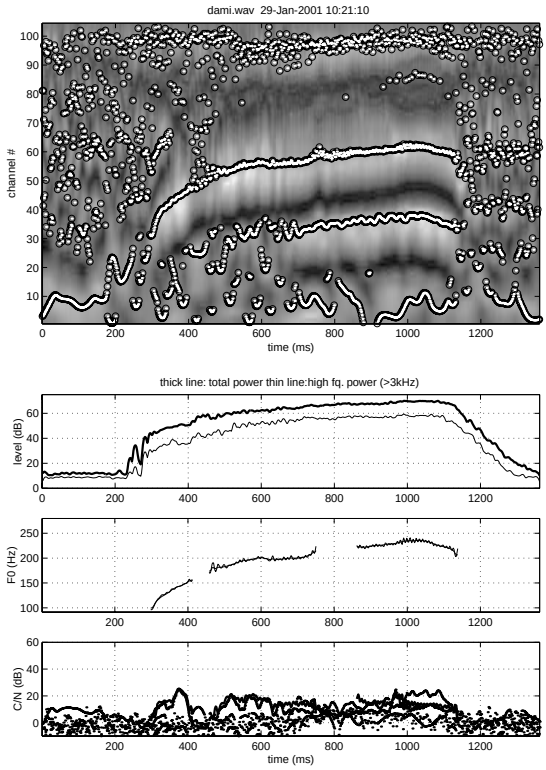

図 6 男性の『だみ声』の音源情報
Fig. 6 Source information for a "dami-goe" produced by a male performer.

求められた軌跡と置き換えることでスケールの歌唱を生成することができる．しかし，4 章で説明した他の要因についてはまったく考慮されていないため，こうして合成された歌唱からは人工的な印象を完全には拭えない．

6.2 特異な発声の STRAIGHT による分析例

ここでは，従来の分析方法ではパラメータ抽出が困難であった特異な発声が，どのように STRAIGHT によって分析され合成されるかを例示する．**図 6** に,『だみ声』の音源情報を STRAIGHT により分析した例を示す．ここでは母音「ァ」を力んで発声することにより収録した，濁って聞こえるサンプルを用いている．最上段の図は，40 Hz から 800 Hz までを 1 オクターブあたり 24 個のフィルタを用いて分析し，正弦波の所在位置を表す不動点を求めた結果を示す．不動点は白抜きの○印で表されている．図の背景の明暗は，それぞれのフィルタ出力に含まれる正弦波成分（carrier）のレベルと背景雑音（noise）のレベルの比として定義される C/N の推定値を示す．ここでは，雑音の少ない部分は明るく表示されている．上から 2 段目は，全帯域のパワーと，3,000 Hz 以上の帯域のパワーを示す．3 段目には，選択された基本周波数の初期推定値と，複数の調波成分を用いて改良された推定値を示す．最下段は，各不動点の C/N の推定値を示す．特殊なフィルタの設計法を用いているため，周期信号を分析すると基本波に対応する正弦波成分以外の C/N は 0 dB 付近となる．また，正常な発声の持続母音の場合，基本波に対応する不動点の C/N は，40 dB 以上となることが多い．

例示した『だみ声』は，220 ms から 1,150 ms あたりまで発声されている．通常の母音であれば，この範囲の基本周波数の軌跡は連続している．しかし，図 6 に示した分析結果では，基本周波数の軌跡が 420 ms 付近と 800 ms 付近で不連続になっている．通常の基本周波数抽出方法では，特異な発声の場合には，このような欠落のある情報が抽出されるだけである．STRAIGHT では，基本周波数だけではなく最上段に示すように不動点と C/N を表すマップという豊かな情報が得られる．このマップには，250 ms 付近から 1,100 ms 付近に至る連続する不動点の系列と，700 ms から 1,100 ms 付近に至る別の連続する不動点の系列が認められる．この表示は，700 ms 付近において声帯振動が二重周期を持つモードに移行していることを意味している．また，最下段のそれぞれの不動点の C/N も，通常の発声の場合に典型的な 40 dB という値と比較すると非常に低く，声帯の振動が異常な状態にあることが分かる．

STRAIGHT による分析結果を用いて，この『だみ声』の再現と，基本周波数や，速度，スペクトル包絡の変換による個人性やスタイルの変換を試みた．予備的な実験の印象では，基本周波数軌跡の欠落する部分でやや人工的な響きが生ずるものの，自然な『だみ声』が得られた．

なお，これらのサンプルと，加工音声の例は，以下のページで参照可能となっている．是非，本資料での説明がどの程度成立しているのかを，自分の耳で確認していただきたい．

`http://www.sys.wakayama-u.ac.jp/~kawahara/scat/`

7. ま と め

望む人の声で，望む感情と表情で，任意の曲の歌唱を合成するという，歌唱合成研究の究極的な目標への 1 つの里程標として，脳のアーキテクチャに倣ったスキャット生成システムの研究構想を提案した．高品質音声分析変換合成系として開発された STRAIGHT を利用して，コーパスベースの合成から出発し，つねに鑑賞に堪える高い品質を維持しながらモデルを導入す

ることで素材を段階的に削減するという戦略は，モデルに対する厳しいベンチマークの機会を提供する．見方を変えれば聴覚情報処理，音声情報処理（特に非言語情報ならびにパラ言語情報処理），音楽情報処理についての総合的ベンチマークとして利用することができるのである．すでに，本論文の前身を研究会で発表した後に，コーパスを用いた歌唱ピッチパターンの自動生成が青野らによって試みられている[40]．具体的な中間目標を設定することによって，このような様々な研究者が競争し協力する場を創り出し，計算機音楽研究の裾野と深さが拡大することを望んでいる．また，その過程を通じて声のレタッチソフトや数多くの表情/表現のプラグインが開発されるとともに，聴覚，音声，音楽への理解が大きく深まることを期待している．

謝辞 本研究は，科学技術振興事業団CREST「脳を創る」領域の『聴覚脳プロジェクト』の支援を受けている．また，一部に科学研究費（基盤C：11650425）の支援を受けた．

参 考 文 献

1) Kawahara, H.: Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisited, *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Muenich, Vol.2, pp.1303–1306 (1997).
2) Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction, *Speech Communication*, Vol.27, No.3-4, pp.187–207 (1999).
3) 河原英紀：聴覚の情景分析が生んだ高品質Vocoder: STRAIGHT，日本音響学会誌，Vol.54, No.7, pp.521–526 (1998).
4) Dudley, H.: Remaking speech, *J. Acoust. Soc. Am.*, Vol.11, No.2, pp.169–177 (1939).
5) Kawahara, H., Katayose, H., de Cheveigné, A. and Patterson, R.D.: Fixed Point Analysis of Frequency to Instantaneous Frequency Mapping for Accurate Estimation of F0 and Periodicity, *Proc. Eurospeech'99*, Vol.6, pp.2781–2784 (1999).
6) 河原英紀，津崎 実，Patterson, R.D.：オールパスフィルタの位相操作による時間構造制御とその知覚への影響について，聴覚研究会資料，H-96-79, pp.1–8 (1996).
7) 誉田雅彰：科学技術振興事業団戦略的基礎研究推進事業（CREST）—発声力学に基づくタスクプラニング機構の構築，日本音響学会誌，Vol.56, No.11, p.771 (2000).
8) 河原英紀：科学技術振興事業団戦略的基礎研究推進事業（CREST）—聴覚の情景分析に基づく音響・音声処理システム，日本音響学会誌，Vol.56, No.11, p.772 (2000).
9) 川人光男：脳の計算理論，産業図書 (1996).
10) Kuhl, P.K., Williams, K.A., Lacerda, F., Stevens, K.N. and Lindblom, B.: Linguistic experience alters phonetic percpetion in infants by 6 months of age, *Science*, No.255, pp.606–608 (1992).
11) McAulay, R.J. and Quatieri, T.F.: Speech Analysis/Synthesis Based on a Sinusoidal Representation, *IEEE Trans. ASSP*, Vol.34, pp.744–754 (1986).
12) 河原英紀，Zolfaghari, P.：不動点に基づく音源情報抽出法の評価について，聴覚研究会資料，H-2000-80 (2000).
13) Titze, I.R.: *Principles of voice production*, Prentice Hall (1994).
14) Fujisaki, H. and Hirose, K.: Analysis of voice fundamental frequency contours for declarative sentences of Japanese, *J. Acoust. Soc. Jpn.* (*E*), Vol.5, No.4, pp.233–242 (1984).
15) Farley, G.R.: A biomechanical laryngeal model of voice F0 and glottal width control, *The Journal of the Acoustical Society of America*, Vol.100, No.1, pp.3794–3812 (1996).
16) Orlikoff, R.F. and Baken, R.J.: Fundamental frequency modulation of the human voice by the heartbeat: Preliminary results and possible mechanisms, *The Journal of the Acoustical Society of America*, Vol.85, No.2, pp.888–893 (1989).
17) Kawahara, H. and Williams, J.C.: Effects of Auditory Feedback on Voice Pitch, *Vocal Fold Physiology*, Davis, P.J. and Fletcher, N.H.(Eds.), chapter 18, pp.263–278, Singular Publishing Group (1996).
18) Larson, C.R., Burnett, T.A., Kiran, S. and Hain, T.C.: Effects of pitch-shift velocity on voice F0 responses, *The Journal of the Acoustical Society of America*, Vol.107, No.1, pp.559–564 (2000).
19) Jones, J.A. and Munhall, K.G.: Perceptual calibration of F0 production: Evidence from feedback perturbation, *The Journal of the Acoustical Society of America*, Vol.108, No.3, pp.1246–1251 (2000).
20) 難波精一郎：聴覚ハンドブック，ナカニシヤ出版 (1984).
21) d'Alessandro, C. and Castellengo, M.: The pitch of short-duration vibrato tones, *The Journal of the Acoustical Society of America*,

Vol.95, No.3, pp.1617–1630 (1994).
22) d'Alessandro, C., Rosset, S. and Rossi, J.-P.: The pitch of short-duration fundamental frequency glissandos, *The Journal of the Acoustical Society of America*, Vol.104, No.4, pp.2339–2348 (1998).
23) 北風裕教，赤木正人：基本周波数の微細変動成分に対する知覚，信学技報，SP99-168 (2000).
24) 本多清志：Biological Mechanisms for Tuning Voice Fundamental Frequency，喉頭，Vol.8, No.2, pp.109–115 (1996).
25) 田中公人，阿部匡伸：基本周波数の変更量に応じてスペクトル包絡を変形する音声合成方式，信学論，Vol.J83-DII, No.8, pp.1722–1732 (2000).
26) Zhu, W. and Kasuya, H.: Perceptual contributions of static and dynamic features of vocal tract characteristics to talker individuality, *IEICE Trans. Fundamentals*, Vol.E81-A, No.2, pp.268–274 (1998).
27) Kuwabara, H. and Takagi, T.: Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method, *Speech Communication*, Vol.10, No.5-6, pp.491–495 (1991).
28) 片寄晴弘：音楽情報処理：第 4 章，文字と音の情報処理，pp.163–224, 岩波書店 (2000).
29) Senjyu, M. and Ohgushi, K.: How are the players ideas conveyed to audience?, *Music Perception*, Vol.4, pp.311–323 (1987).
30) 河原英紀：声を使って聴覚を探る，日本音響学会誌，Vol.53, No.9, pp.731–737 (1997).
31) 矢田部学，遠藤康男，粕谷英樹：歌声の基本周波数の動特性，音講論，pp.383–384 (1998).
32) 小田切わか菜，森　大毅，粕谷英樹：歌声のピッチ遷移に関する検討，音講論，pp.537–538 (2000).
33) Kawato, M.: Internal models for motor control and trajectory planning, *Current Opinion in Neurobiology*, Vol.9, No.6, pp.718–727 (1999).
34) Kawato, M., Furukawa, K. and Suzuki, R.: Hierarchical neural-network model for control and learning of voluntary movement, *Biological Cybernetics*, Vol.57, pp.169–185 (1987).
35) 河原英紀，Zolfaghari, P., de Cheveigné, A., Patterson, R.D.：周波数から瞬時周波数への写像の不動点を用いた音源情報の抽出について，信学技報，SP99-40 (1999).
36) 河原英紀，阿竹義徳：音声の群遅延特性に基づく声門閉止等のイベント抽出について，信学技報，SP99-171 (2000).
37) Kawahara, H., Atake, Y. and Zolfaghari, P.: Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay, *Proc. ICSLP'2000*, Beijin, China, pp.664–667 (2000).
38) 河原英紀，Zolfaghari, P.：群遅延情報を利用した音声の駆動情報の多重解像度分析について，信学技報，EA2000-35, pp.63–70 (2000).
39) Kawahara, H., Estill, J. and Fujimura, O.: Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT, *Proc. 2nd MAVEBA*, Firenze, Italy (2001).
40) 青野裕司，水野秀之，阿部匡伸：コーパスを用いた歌唱ピッチパターンの自動生成，音響学会秋季講演論文集，Vol.I, pp.225–226 (2001).

(平成 13 年 6 月 18 日受付)
(平成 13 年 12 月 18 日採録)

河原　英紀（正会員）

昭和 47 年北海道大学工学部電子工学科卒業．昭和 52 年同大学院博士課程修了．工学博士．平成 9 年度より，和歌山大学システム工学部教授．聴覚メディア処理，音声分析変換合成，聴覚情景分析モデルの研究に従事．平成 9 年度日本音響学会佐藤論文賞，EURASIP Best Paper Award 1998/99 受賞．電子情報通信学会，日本音響学会，神経回路学会，認知科学会，米国音響学会，IEEE 各会員．科学技術振興事業団戦略的基礎研究推進事業（CREST）『脳を創る』領域『聴覚脳プロジェクト』研究代表者．http://www.sys.wakayama-u.ac.jp/~kawahara

片寄　晴弘（正会員）

昭和 61 年大阪大学基礎工学部制御工学科卒業．平成 3 年同大学院博士課程修了．工学博士．平成 9 年度より，和歌山大学システム工学部助教授．音楽情報処理，感性情報処理，インタラクティブアート制作の研究に従事．平成 2 年情報処理学会学術奨励賞受賞．電子情報通信学会，人工知能学会，ICMA 各会員．科学技術振興事業団さきがけ 21 研究員．http://www.sys.wakayama-u.ac.jp/~katayose/