音声研究　第13巻第1号
2009(平成21)年4月
29–39頁

Journal of the Phonetic Society
of Japan, Vol.13 No.1
April 2009, pp.29–39

TANDEM-STRAIGHTと音声モーフィング：感情音声と歌唱研究への応用

河原　英紀*・森勢　将雅**

TANDEM-STRAIGHT and Voice Morphing: Applications to Emotional Speech and Singing Research

Hideki KAWAHARA* and Masanori MORISE**

SUMMARY: A versatile speech analysis, modification and resynthesis system TANDEM-STRAIGHT and voice morphing based on it are introduced in relation to classical ideas starting from VOCODER in 1939. Interactive graphical user interfaces for direct manipulation of parameters and morphing parameters are also introduced to provide hints for readers to find make use of the procedures. Also an application of morphing procedure for detecting a high level auditory after effects in gender perception are introduced to highlight a prospective strategy for using morphing as a microscope in perceptual research.

キーワード：音声分析合成，VOCODER，音声変換，モーフィング，パラ言語情報，インタフェース

1. はじめに

携帯メールなど，絵文字や文字アートを使って，テキストだけによるよりも豊かなコミュニケーションが試みられている。しかし，心に直接届くメッセージの媒体として，声にまさるものは（今のところ）ない。音楽も，歌声が加わることでより心を動かすものとなる。2007年に発売された，歌声を簡単に合成できる「初音ミク」[1]の驚異的な売り上げは，他の要因も大きかったにせよ，魅力的な歌声があってこその記録であろう。発売直後から，動画投稿サイトには「初音ミク」の歌声が溢れ，数ヶ月で百万回以上再生される曲も生まれている。新しい音声の加工技術TANDEM-STRAIGHT（Kawahara et al. 2008）とモーフィングは，このような愉しみの可能性をさらに広げるだけではなく，声をテキスト以上のものにしている魅力の源を研究するための新しい手段を提供する。

2. 温故知新

新しい音声の加工技術TANDEM-STRAIGHT（およびその前身のSTRAIGHT（Kawahara et al. 1999））に使われているアイデアや技術は，必ずしも新しいものだけではない。そのルーツの幾つかは，昔の論文の中に見つけることができる。それらの昔のアイデアが，近年の情報処理能力・容量の爆発的な進歩（30年間で百万倍）により，新しい意味を持ち始めている。この原稿を執筆しているラップトップには，1929年創刊の米国音響学会誌全巻など，幾つかの学会誌の数十年分がインストールされている。ネットでの検索と併せると，アイデアのルーツを訪ねるための，時間や場所の制約が消えつつある。TANDEM-STRAIGHTのアイデアの多くも，源をそれらの資料の中に見つけることができる。ここでは，まず，STRAIGHTの構造の基礎となっている，Bell研究所のDudleyにより発明された世界初の電気的音声合成器VODERとVOCODER(Dudley 1939)から始めることとしたい。

* 和歌山大学（Wakayama University）
** 立命館大学（Ritsumeikan University）

2.1 VOCODER 再訪

VODER をネットで検索すると，1939 年のニューヨーク万博での巨大な展示の写真がすぐに見つかる。アナウンサーの言葉を復唱する VODER の声を聴くこともできる。コンピュータが出現する以前のこの音声合成器は，オペレータの女性が「演奏」していた。Smithsonian 博物館の web[2] では，演奏に用いられた鍵盤型のインタフェースの写真を見ることができる。

言葉を正確に聴き取ってもらうためには，一秒間に 4000 回程度の頻度で振動する声の成分を伝える必要がある。しかし，VODER を演奏する指の動く頻度は，一秒間に数十回を超えることはない。それにもかかわらず，VODER の言葉は明瞭に理解できる。これは，音声信号の性質が信号そのものの速度よりも遥かにゆっくりと変化していることを意味する。人間のオペレータの代わりに，指と同じ役割を持つ信号を音声波形から取り出す電気回路を置いたものは，VOCODER と呼ばれた。

明瞭ではあるものの，VODER の声の品質は良くない。VOCODER も，品質の悪い合成音声の代名詞として使われていた。しかし，実際の発声器官を機能的に見た仕組みは，VODER と違ってはいない。つまり，VODER や VOCODER の音が悪い原因は，仕組みにではなく，動かし方と，仕組みの調整にあると考えた方が良い。STRAIGHT そして TANDEM-STRAIGHT では，問題となっていたそれらを精密化した。文字通り，現代版の VOCODER である。

鍵は，パワースペクトルの推定にある。母音のような周期性のある音のスペクトログラムには，周期性の影響により，繰り返しのある構造が，時間方向あるいは周波数方向に表れる。図 1 に示す，広帯域のスペクトログラムと狭帯域のスペクトログラムの中間にあたる条件で求めたスペクトログラムの例では，周波数と時間の両方に周期性の影響が表れている。TANDEM-STRAIGHT では，このスペクトログラムから周期性の影響だけを取り除いて図 2 に示すような滑らかなスペクトログラム（以下，STRAIGHT スペクトログラム）が求

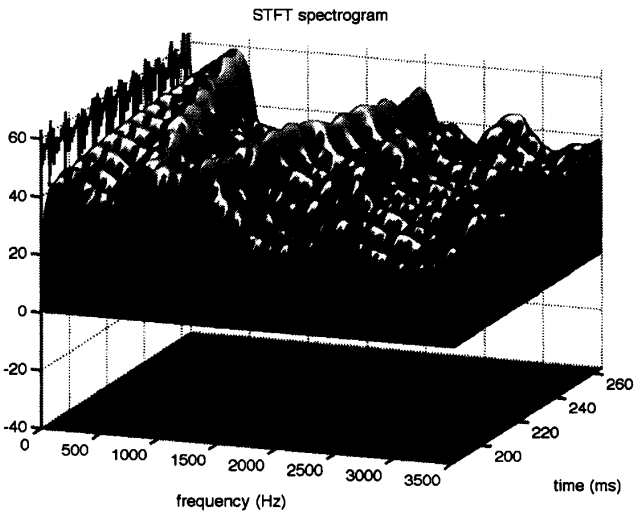

図 1　通常の短時間 Fourier 変換により求めたスペクトログラムの 3 次元表示。試料は男声話者の連続発声した「あいうえお」の一部。下には，同じものの 2 次元表示を示す。

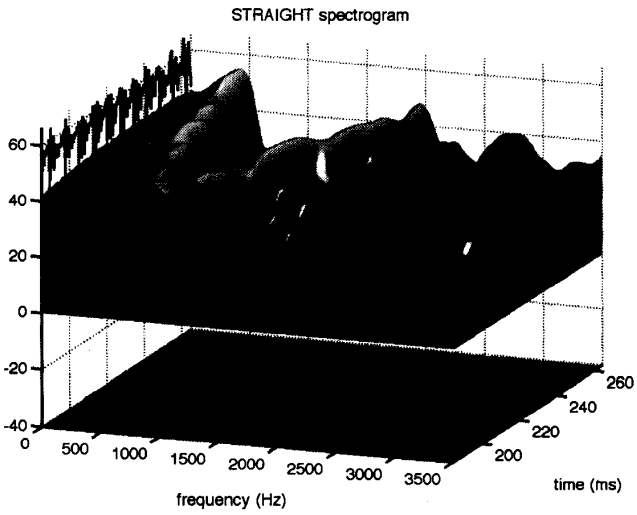

図 2　TANDEM-STRAIGHT により求めたスペクトログラムの 3 次元表示。試料と表示法は図 1 と同じ。

められる。この滑らかな表現が，実用的に使える品質での音声モーフィングなどの加工を可能にした TANDEM-STRAIGHT の基礎になっている。この表現は，パワースペクトルの計算の見直しと，標本化定理というデジタル信号処理のルーツの見直しから生まれた。それらを説明する以下の二つの節は，やや専門的になる。

2.2 パワースペクトル推定再訪

1965 年の高速 Fourier 変換（FFT）の発明は，

デジタル信号処理の応用範囲を一挙に拡大した。FFT が無ければ，携帯電話の実用化も数十年の範囲で遅れたに違いない。1960 年代の後半には，パワースペクトルなどの基本的な量を，FFT を用いて計算するアルゴリズムが次々と発表された。そのパワースペクトルの計算法として現在でも広く用いられている Welch の方法（Welch 1967）では，重畳した複数の区間に分割した信号それぞれのパワースペクトルの平均として推定値が求められる。こうすることにより，信号の統計的な揺らぎの影響が軽減される。ところで，分析対象とする信号が周期信号の場合には，基本周期の半分だけ位置の異なる二つの区間で計算したパワースペクトルを平均するという簡単な方法で，周期性に起因する時間的変動を完全に無くすることができる。周期信号の場合にだけ成立するこの性質を利用して発明された TANDEM（森勢ほか 2007）は，上記の Welch 法の特殊な場合に相当する。図 1 にあった周期性による時間方向の変動は，この TANDEM を用いることで，完全に取り除かれる。残っている周波数方向の周期性の影響の除去では，標本化定理の新しい見方が重要な役割を果たす。

2.3 標本化定理再訪

音声のように連続的に変化するアナログ信号を計算機等で処理するためには，離散的な時刻での値を取り出す標本化が必要になる。標本化されたデジタル信号と元のアナログ信号が同じになる条件は，1949 年に，Shannon と染谷がそれぞれ独立に標本化定理として明らかにしていた（Shannon 1949，染谷 1949）。CD や DVD，ネット上の音声，画像など，デジタルメディアのほとんど全ては，この標本化定理に基づいている。実は，前の節で残されていた，調波構造を持つスペクトルから周期性の影響を取り除くという課題は，この標本化定理が対象としている課題と同じ構造を持つ。ただし，ホルマントの近くのスペクトル形状には，この標本化定理が要求する条件を満たしていないという問題がある。

この問題の解決策は，証明から半世紀を経て，再び研究が活発に行われている標本化定理の新しい見方（Unser 2000）からもたらされた。この新しい見方では，形状の完全な復元を要求するのではなく，調波位置での値の復元のみを要求する。コンシステント再構成と呼ばれるこの見方に基づくことで，最初の STRAIGHT（Kawahara et al. 1999）と比較すると遥かに処理量が少なく理論的にも見通しの良い計算法が発明された（Kawahara et al. 2008）。図 1 から図 2 を求める計算には，TANDEM と併せてこの方法が用いられている。

3. 音声の加工とモーフィング

TANDEM-STRAIGHT は，分析部と合成部から構成される。分析部では，これまでに説明して来た STRAIGHT スペクトログラムと，基本周波数が求められる他，有声摩擦音のような周期成分と非周期成分が混合した状態を表すための非周期性指標が求められる。こうして求められた三種類のパラメタを加工し合成部に渡すことで，様々に加工された音声を合成することができる。

3.1 パラメタの加工と加工音声の合成

最初の STRAIGHT（Kawahara et al. 1999）の実装には，科学技術計算用環境である Matlab[3] が用いられていた。Matlab 環境下では，システムが提供する機能を用いるだけで対話的にデータを表示・加工することができる。また，C 等の言語を用いる場合と比較すると，Matlab を用いたプログラムは 1/10 以下の行数で記述出来る簡単なものとなる。そのため，最初の STRAIGHT では，デモ用の簡単なユーザインタフェースを用意しただけで，応用に必要なプログラミングは，利用者がそれぞれの目的に応じて開発することを前提としていた。このように利用者とってやや敷居の高いシステムであったにも関わらず，STRAIGHT は，音声知覚の様々な問題に応用されてきた（内田 2000，Liu 2004，Smith et al. 2005）。これらの例の多くでは，分解されたそれぞれのパラメタを，

自然性を大きく低下させることなく独立に操作できるという特徴が生かされている。他の応用事例や詳しい紹介などは，記述は最初の STRAIGHT に限られているが，展望論文（河原 2007）を参照されたい。

3.1.1 パラメタ加工用対話的インタフェース

最初の STRAIGHT では処理速度が非常に遅く必要なメモリ量が膨大になるため，対話的なインタフェースを提供することは実際的ではなかった。TANDEM-STRAIGHT の発明は，この速度とメモリ量の問題を解決し，対話的にパラメタを加工するためのインタフェースの開発を促した（河原ほか 2009）。なお，TANDEM-STRAIGHT においても開発効率と対話的環境を重視して Matlab が用いられている。

図 3 に，新たに開発されたインタフェースを示す。最上段には，STRAIGHT スペクトログラムが表示され，その下に 5 種類の操作パネルが表示されている。図 3 では，基本周波数の操作パネルが拡大表示され，他は縮小表示されている。操作パネルとしては，他に，パワー，周波数軸の伸縮，時間軸の伸縮，非線形指標のためのものが用意されている。必要に応じてこれらの操作パネル内をクリックすることにより，そのパネルを拡大表示することができる。拡大表示されたそれぞれの操作パネル内では，軌跡の全体あるいは一部を選択して平行移動させることと，軌跡そのものを手書きして修正することができる。図 3 では，灰色（画面では緑色）の太線の部分が，手書きされた軌跡を表している。なお，手書きされた軌跡は，指定

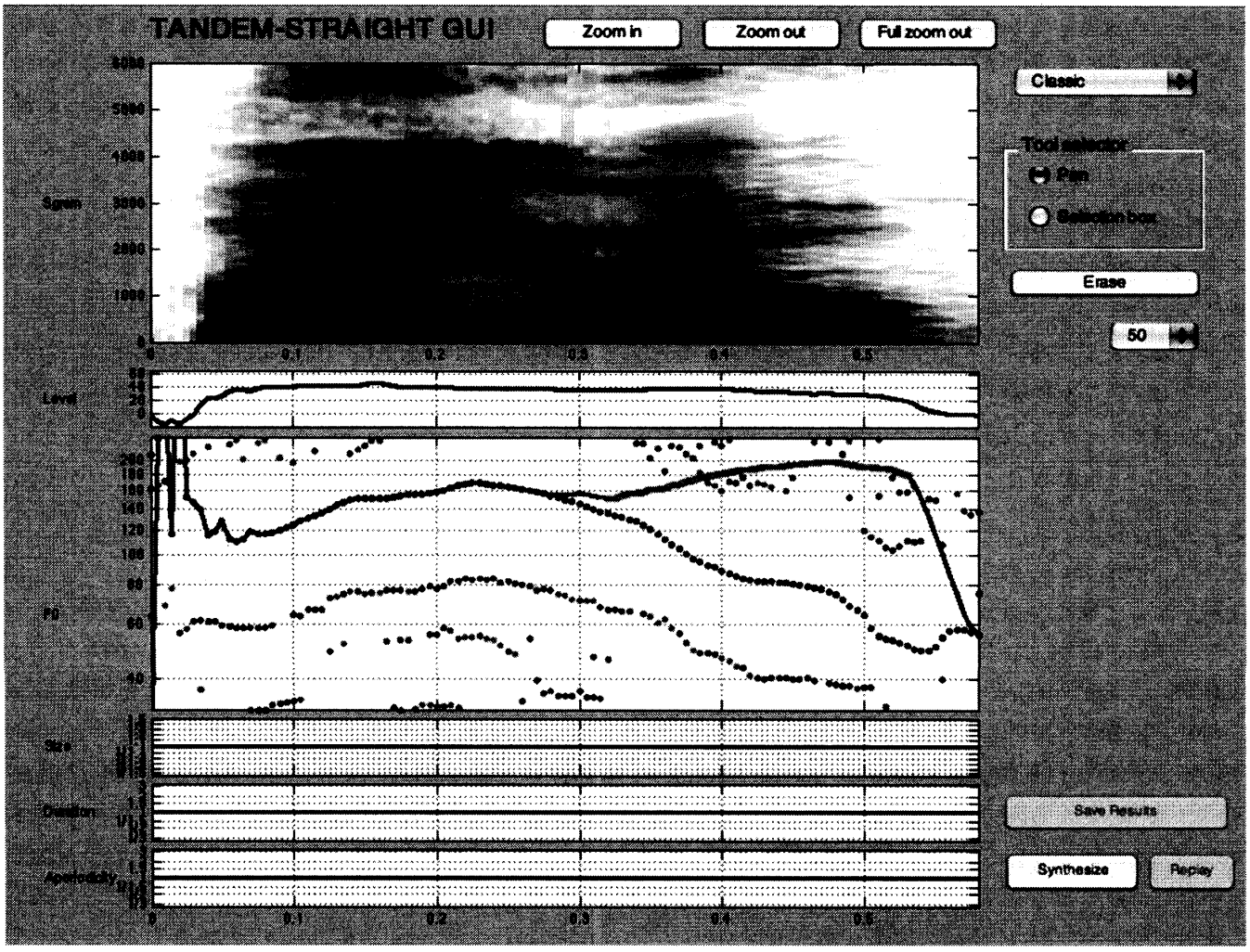

図 3　STRAIGHT で求められたパラメタの加工に用いるインタフェース。

された遷移幅（この例では50ms）の滑らかな遷移部分を介して元の軌跡に移行する。

このような操作で加工されたパラメタから，右下のボタンをクリックすることにより，加工された音声が合成される。合成された音声は，音声合成用のボタンの上にあるボタンをクリックすることにより，パラメタ加工の情報と併せて，ファイルに書き出される。ユーザは，プログラムを書くことにより，これらの情報および分析部と合成部の関数を直接利用することもできる。

しかし，このような加工だけで希望通りの効果を得ることは難しい。例えば，早口の発話をゆっくりとしたものに変える場合には，一様に時間軸を引き伸すのではなく，母音定常部だけを引き伸すようにしないと，病的な印象の音声になってしまう。声を高くする場合も，基本周波数を上昇させるだけではなく，同時に周波数軸を引き伸して体のサイズをある程度小さくしないと，不自然な声になってしまう。普通に話された音声を感情の込められた音声に変えることは，更に難易度が高く，このような加工だけではほとんど不可能に近い。次に説明する音声モーフィング（Kawahara and Matsui 2003）は，これらの困難な加工のための別のアプローチを提供する。

3.2 モーフィングによる音声加工

モーフィングは，与えられた二つの事例から，それらの事例の中間にあたるものを作り出す。音声モーフィングを用いると，怒りの感情のこもった音声試料と，悲しみの感情のこもった音声試料から，例えば，それらを7対3の割合で混合した音声を合成することができる。

モーフィングでは，悲しみや怒りの感情がどのような音声のパラメタに対応しているかを予め理解している必要はない。目的とする感情が十分に表現されている音声試料をまず用意し，それらの試料をSTRAIGHTにより分析して求めたパラメタを適当な割合で補間するだけで，様々な中間の音声を合成することができる。見方を変えると，この補間の操作の中に，二つの試料の間の知覚的属性の違いを与える物理的要因が全て含まれていることになる。補間の割合（モーフィング率）は，パラメタ毎に違っても良いし，モーフィング率が時間とともに変化しても構わない（Kawahara et al. 2009）。この自由度により，二つの試料の間を補間する経路を，幾つでも作り出すことができる。

以下では，まず音声モーフィングの応用例を幾つか示し，次に，それらを作成するためのインタフェースを紹介する。

3.2.1 感情音声のモーフィング

STRAIGHTに基づく音声モーフィングは，最初に感情音声の研究に応用された（Matsui and Kawahara 2003, Kakehi et al. 2008）。モーフィングされた感情音声の主観評価結果を多次元尺度構成法で分析することにより，「喜び」「怒り」「哀しみ」が「平静」を重心に置く三角形の各頂点に配置され，モーフィングされた音声がそれぞれの感情間を結ぶ線上に配置されることが示された。この知見に基づいて，図4に示したインタラクティブな作品[4]が制作され，日本科学未来館の特別企画展[5]で展示された。この作品では，声優により「喜び」「怒り」「哀しみ」の感情を込めて録音された音声（STRAIGHTによる再合成音）が三角形の頂点に配置されている。モーフィングにより合成された音声は，頂点同士と重心を結ぶ線上に配置されており，マウスのクリックにより当該位置に対応する音声が再生される。重心にある「平静」は，「喜び」「怒り」「哀しみ」の三種類の感情音声の平均値を用いて合成されている。

3.2.2 歌唱のモーフィング

音声モーフィングは，歌唱音声の研究にも応用されている。曲げセンサーや圧力センサーを組込んだ指人形を使い，歌唱の表現を直感的に操作するインタフェースが提案されている（Yonezawa et al. 2005）。このインタフェースでは，ユーザのジェスチャーを歌唱の表現の滑らかな変化に対応づけるために，音声モーフィングが用いられている。主観評価実験の結果によれば，物理パラメタの操作量と知覚される演奏表現の変化量の対応関係は，素材として用いた事例の近くでは飽和する。

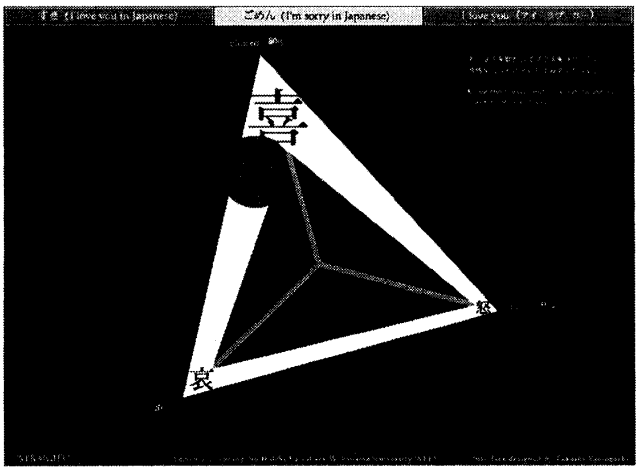

図4　感情音声モーフィングを利用したインタラクティブな展示作品。この例では，喜びと哀しみが8対2の割合で混合されている。（デザイン：山口崇，日本科学未来館特別企画展[5]より）

この対応関係を変換関数を用いて線形化することにより，より自然な操作感を得ることができることが報告されている。

モーフィングに関わる複数のパラメタと変換を組み合わせることで，歌い回しや声質など，直感的に把握し易い属性を直接操作するための研究も行われている（豊田ほか 2006，河原ほか 2007）。この実験では，基本周波数と時間軸の変換が歌い回しに，スペクトルのレベルと非周期性および周波数軸の変換が声質に強く関連していることと，歌い手についての総合的な判断が主に声質により行われることが報告されている。また，これらの結果を用いて，リアルタイムに歌唱音声の声質と歌い回しを操作するインタフェースが提案されている（森勢ほか 2008）。

図5に，提案されたインタフェース V.morish の操作画面を示す。左側がリアルタイムに対話的に声質と歌い回しを操作する部分であり，操作の履歴が徐々に透明になる擬似的な3次元の軌跡により表示されている。右側の二つのパネルは，声質と歌い回しの軌跡をオフラインで編集するために用いられる。この V.morish の実装では，予め様々なモーフィング率の音声を合成し，インタフェースの操作に応じて出力する合成音声を切り替えるという方法が用いられていた。しかし，時間軸の

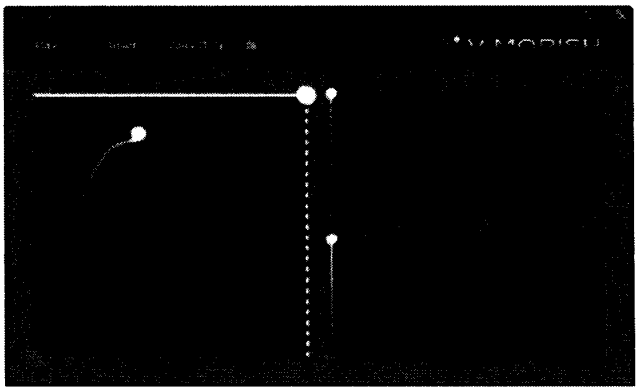

図5　リアルタイムに歌い回しと声質を変換するインタフェース V.morish の操作画面。

操作を正しい形で操作対象に含めることはこの方法では困難であり，また，操作量が外挿処理を要求するときには場合により加工が破綻する。

これらの問題を解決するために，様々な属性に対応するパラメタのモーフィング率が時間とともに独立に変化する場合への拡張を含み，外挿処理でも破綻を生じない方法が，TANDEM-STRAIGHT に基づいて新たに定式化された（河原ほか 2008，Kawahara et al. 2009）。以下に紹介するモーフィング用のインタフェースは，この新しい定式化に基づいている。

3.2.3　モーフィング用インタフェース

前の節で紹介したようなモーフィング音声を合成するためには，下ごしらえが必要となる。まず，それぞれの試料の STRAIGHT スペクトログラム上に，重ね合わせるための基準となる点（基準点）を設定することが必要になる。基準点は，位置のずれが合成された音声品質の劣化につながるような特徴的な点に設定する。例えば，母音や子音の開始や終了時点，ホルマント周波数の遷移開始および終了位置などが該当する。基準点の個数は，時間方向には，含まれる音素の個数の2倍程度，周波数方向には3～5個が目安となる。

この方法では，数秒の音声でも設定すべき点の数は100個を超える。設定には，音声生成過程と音声知覚およびデジタル信号処理の基礎的な知識が必要であり，試行錯誤による調整も必要になる。このコストのかかる手作業を排除するために，コ

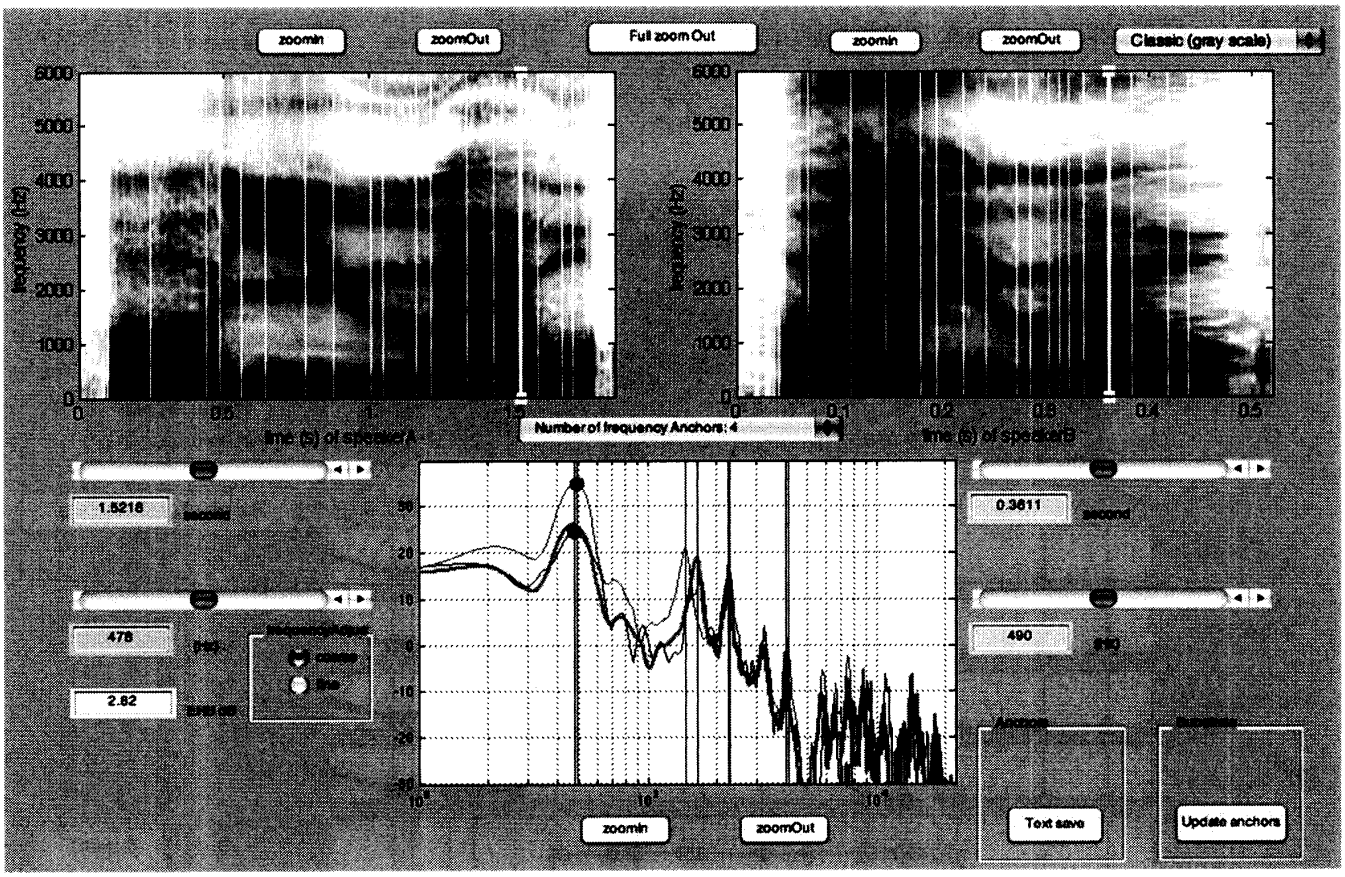

図 6　モーフィングの基準点設定のためのインタフェース。

ンテンツ制作などの最終的な加工音声だけを必要とする用途に向けて，モーフィングの自動化の研究が進められている（Onishi et al. 2008）。なお，予備検討のように品質がそこそこであっても手早く結果が必要な場合には，もう一つの抜け道がある。同性の話者の音声同士の場合には，音素境界の位置に基準点を置くだけで周波数方向には基準点を置かない簡易モーフィングにより，比較的良好な品質の音声を合成することができる（西田ほか 2008）。

しかし，音声知覚研究のための刺激を作成するような場合には，基準点をどこに置くか自体が重要な実験条件となるため，これらの省力化手段を利用することはできない。このような場合に必要となる多数の基準点の設定作業を容易にするために，図 6 に示すインタフェースが用意されている。この例では，左側にゆっくりと話した連続母音「あいうえお」の STRAIGHT スペクトログラムが表示され，右側に早口で話したものが表示されている。

図中の縦線は，基準点の置かれている時刻を表す。太くなっている縦線は，現在作業を行っている時刻を示しており，その時刻における STRAIGHT スペクトログラムの断面が，インタフェースの下のパネルに表示されている。これらの断面のグラフは，左右いずれの試料に対応しているかが容易に把握できるよう，色分けされている（本資料ではグレースケールに置換されている）。グラフの中の縦線は，周波数方向の基準点の置かれている位置を示す。また，灰色（画面では緑色）の太線は，設定された基準点を用いて変換されたスペクトルを表す。基準点の位置は，この変換されたスペクトルが目標となるスペクトルにできるだけ重なることを目標として調整される。なお，それぞれのパネルではズームと表示対象のドラッグによる視点の移動が可能になっている。

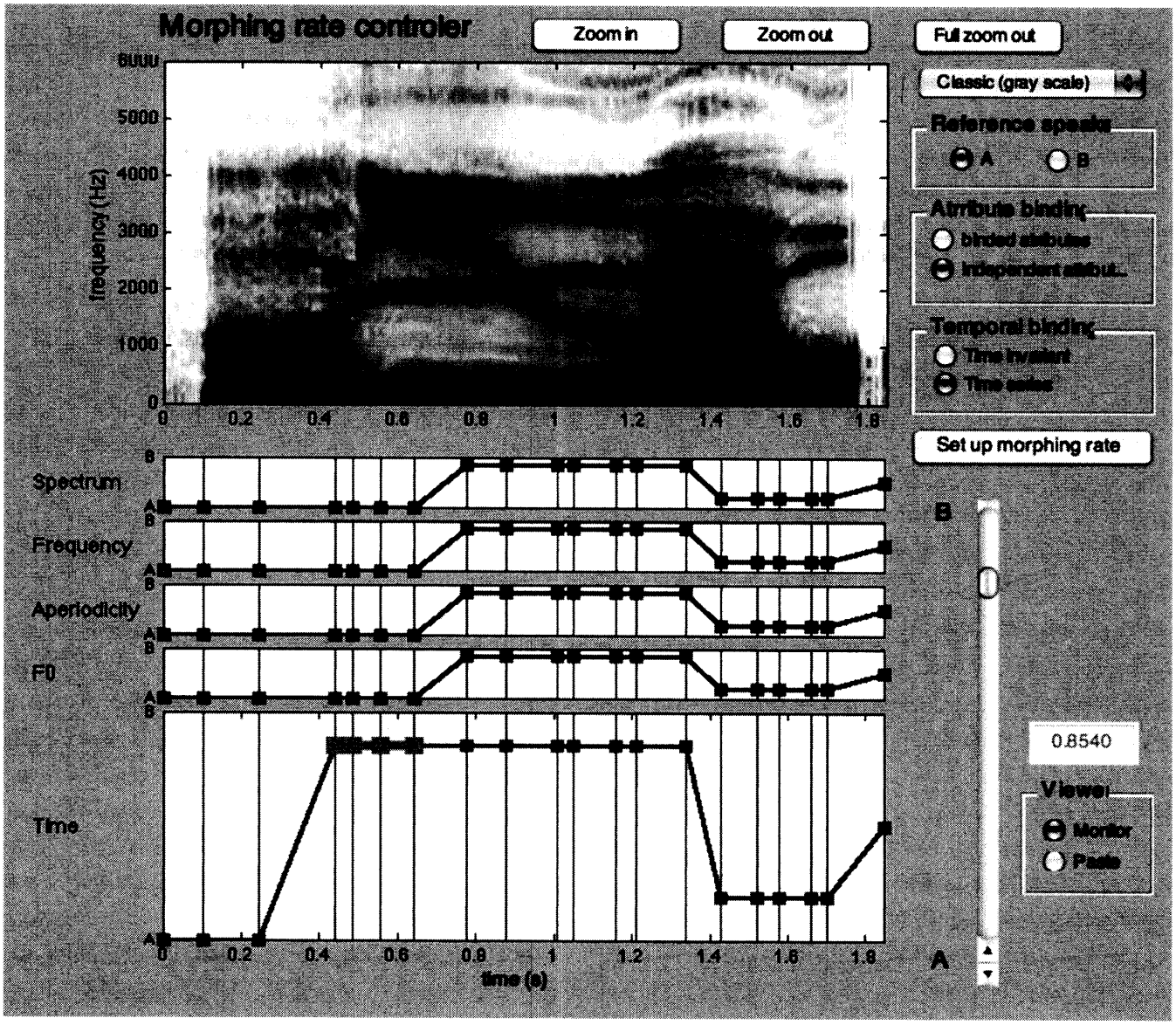

図7　属性毎のモーフィング率時系列設定のためのインタフェース。

複数の属性のモーフィング率の時系列を設定する場合にも，多くの作業が必要となる。この作業を支援するために図7に示すインタフェースが用意されている。パラメタ加工用のインタフェースと同様に，操作パネルはパネル内のクリックにより拡大表示される。必要に応じて時間方向と属性間の操作量をバインドして一括して調整することで，多数の操作点での値を効率よく設定することができる。図では，時間および属性の双方のバインドが解除されており，時間属性のパネル中の灰色（画面では緑色）の部分が，操作のために選択されている。これらの操作を文章だけで把握することは難しい。STRAIGHTを紹介するページ[4]に実際の操作の様子を収録したムービーをリンクしてあるので，それらを参照されたい。

3.2.4　対話的インタフェースの功罪と対策

対話的に音声のパラメタやモーフィング率を操作して，その結果を直ぐに音として確認できる環境は，探索的研究の効率を大きく向上させる。また，実験用の刺激作成に含まれる定型的作業の効率を向上させることで，より複雑で巧妙な刺激の作成を間接的に促進する。しかし，対話的インタフェースには，操作を実装するための様々な既定値が組込まれており，プログラム開発者の想定し

ている操作に発想を制約されてしまう危険もある。また，探索的手法を用いる場合，様々な試行錯誤を行うこととその経過と結果を正確に記録することとは，往々にして両立しない。

これらの問題を軽減するために，今回のシステムでは，インタフェースとプログラミングのそれぞれに以下のような対策を施している。まず，対話的インタフェースには，試行錯誤の結果を再現出来るだけの補足情報を構造体として記録する機能を設けている。プログラムは，対話と分析・合成などの信号処理とを明確に分離した構造とし，同じ形式の構造体を用いて関数を呼び出すだけで簡単に加工音声を合成できる仕様としている。プログラミングのスタイルも，以前の STRAIGHT とは大きく変え，数十文字に及ぶ長い変数名と構造体を多用することで理解し易いものとしている。

4. モーフィングの拓く可能性

爆発的な計算能力の向上と膨大な音声データの蓄積を背景として，統計的モデルに基づく音声処理技術は着実に進歩し，テキストからの音声合成や話者変換は，高い品質と柔軟性を併せ持つに至っている（Zen et al. 2007, Toda et al. 2007）。しかし，それらの進歩は，声の個性や感情，パラ言語情報に分類される声による多様な表情・表現の深い理解には，必ずしも結びついてはいない。この間隙を埋め，知覚とその物理的基盤を深く理解するためには，収録された音声資料の統計的分析と併せて，探索的手法による仮説形成と加工音声を用いた主観評価実験による検証を進める必要がある。ツールとしての TANDEM-STRAIGHT とそれが可能にした，パラメタの定量的な操作と合成音声の高い自然性を両立させるモーフィング技術は，このような研究を進めるための手段となる。

STRAIGHT のパラメタにより表現された二つの音声試料の間をモーフィングで結ぶ経路には，試料の知覚的属性を変換する上で必要な情報が全て含まれている。この経路は，それぞれの試料の STRAIGHT のパラメタと変換関数およびモーフィング率により客観的に記述することができ，しかも，自由に値を操作することができる。この経路を利用して，物理的に加えた操作量と，合成された音声の知覚的な属性の変化から，知覚的属性の物理的基盤を調べるための手掛かりを得ることができる。ただし，こうして求められる知見は，個別の試料に依存したものである。これら個別の多数の知見の中から，いかにして知覚属性の一般的な物理的基盤を見つけ出すかが研究の重要な課題となる。母音情報に基づく音声変換（Onishi et al. 2007）や，話し声から歌声への変換システム（齋藤ほか 2008）は，それぞれ音声知覚における話者適応性の知見（加藤・筧 1996）や，歌唱音声の物理的特徴の分析と知覚実験の知見に基づいて，そのような一般的な変換法を見出そうとする試みの例である。

4.1 知覚研究の顕微鏡

モーフィングには，もう一つの興味深い利用法がある。モーフィングにより刺激連続体を作り，それを物差しとして知覚の残効を調べることで，背景にある処理モジュールを調べようとする方法である。ここでは，男声と女声の間のモーフィング音声を用いて，声の性別判定で残効が生ずることを示した研究（Schweinberger et al. 2008）について紹介する。

この実験では，ドイツ人男女各 5 名の発声した 4 種類の VCV 音節（/aba/, /aga/, /ibi/, /igi/）を 44,100Hz, 16bit で収録した音声が用いられた。録音された音声は，886ms の長さに揃えられた後，7 段階のモーフィング率（80%, 70%, 60%, 50%, 40%, 30%, 20%）のモーフィング音声が作成された。5 組の男声と女声のペアの中の一組は，練習試行用に用いられ，残りの 4 組のペアからの合計 112 個の音声が実験に用いられた。

男声，女声あるいは 50% モーフィングの中声の音声刺激 4 個からなる順応刺激が呈示された後に，500ms の間隔を空けて試験対象となる刺激が呈示され，発声者の性別の判断が求められた。こ

の実験と併せて，男性と女性の発話映像や，基本周波数を合わせた正弦波による刺激，男性的な名前や女性的な名前を順応刺激として，同様な実験が行われた。被験者には，二択でできるだけ速く回答することが求められた。12名～24名の被験者による回答から，モーフィング率と「女声」の判定との関係が求められ，主観的等価点（PSE）が求められた。その結果，順応刺激に音声を用いた場合に，順応刺激と逆方向にPSEが移動することが明瞭に認められた。ここでは，モーフィングによる刺激連続体の利用が，直接測定することの困難な影響を，PSEの移動として定量的に測定することを可能にした。ここでは，モーフィング音声は，いわば，知覚研究における顕微鏡のような役割を果たしている。一つの有用な研究手法と考えられる。

5. おわりに

VOCODERの現代版である音声分析変換合成技術であるSTRAIGHTは，その基礎となるアルゴリズムが理論的な見通しと効率が良い新しいものにより完全に置き換えられてTANDEM-STRAIGHTとなった。本稿は，現在も活発に改良が続けられているTANDEM-STRAIGHTの現状と，それに基づく有望な応用技術である，多様な属性を時系列により制御出来る拡張されたモーフィングについて紹介した。また，これらのために新たに用意した対話的インタフェースと応用事例について紹介することにより，読者の研究への応用を考える際のヒントを提供することを試みた。

謝　辞

TANDEM-STRAIGHTおよび多属性時変モーフィングの研究は，現在，科学研究費補助金基盤（A）19200017と科学技術振興機構による戦略的創造研究推進事業のデジタルメディア領域CrestMuseプロジェクトのの支援を受けて進められている。

〔注〕

1) 初音ミク（2007）キャラクターボーカルシリーズ，クリプトン・フューチャー・メディア株式会社．
2) http://scienceservice.si.edu/ の中の検索で，“voder”を入力する。
3) http://www.mathworks.com/products/matlab/
4) http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_j.html
5) 日本科学未来館特別企画展（2005）『恋愛物語展—どうして一人ではいられないの？—』4/24～8/15.

参考文献

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H. (2008) “TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-free Spectrum, F0, and Aperiodicity Estimation,” *Proc. ICASSP2008*, Las Vegas, 3933–3936.

Kawahara, H., Masuda-Katsuse, I. and de Cheveigné, A. (1999) “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a reptitive structure in sounds,” *Speech Communication*, 27, 187–207.

Dudley, H. (1939) “Remaking speech,” *Journal of the Acoustical Society of America*, 11, 169–177.

Welch, P. D. (1967) “The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms,” *IEEE Trans. Audio and Electroacoustics*, AU-15, 70–73.

森勢将雅，高橋徹，河原英紀，入野俊夫（2007）「窓関数による分析時刻の影響を受けにくい周期信号のパワースペクトル推定法」『電子情報通信学会誌 D』90-D: 3265–3267.

Shannon, C. E. (1949) “Communication in the presence of noise,” *Proceedings of the IRE*, 37, 10–21.

染谷勲（1949）『波形伝送』東京：修教社.

Unser, M. (2000) “Sampling–50 years after Shannon,” *Proceedings of IEEE*, 88, 569–587.

内田照久（2000）「音声の発話速度の制御がピッチ感及び話者の性格印象に与える影響」『日本音響学会誌』56，396–405.

Liu, C. and Kewley-Port, D. (2004) “Vowel formant discrimination for high-fidelity speech”, *J. Acoust. Soc. Am.*, 116, 1224–1233.

Smith, D., Patterson, R., Turner, R., et.al. (2005) "The processing and perception of size information in speech sounds", *J. Acoust. Soc. Am.*, 117, 305–318.

河原英紀（2007）「Vocoderのもう一つの可能性を探る—音声分析変換合成システムSTRAIGHTの背景と展開—」『日本音響学会誌』63，442–449.

Kawahara, H. and Matsui, H. (2003) "Auditory Morphing based on an Elastic Perceptual Distance Metric in an Interference-free Time-frequency Representation," *Proc. ICASSP'03*, Hong Kong, 256–259.

Kawahara, H., Morise, M., Takahashi, T., et.al. (2009) "Temporally variable multiaspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *Proc. ICASSP'09*, Taipei. (accepted for publication)

Matsui, H. and Kawahara, H. (2003) "Investigation of Emotionally Morphed Speech Perception and its Structure Using a High Quality Speech Manipulation System," *Proc. Eurospeech'03*, Geneva, 3157–3160.

Kakehi, K., Sogabe, Y. and Kawahara, H. (2008) "Research on emotinal perception of voices based on a morphing method." In Izdebski K. (ed.) *Emotions in the human voice*. (pp.1–14). SanDiego: Plural Publishing.

Yonezawa, T., Suzuki, N., Mase, K. and Kogure, K. (2005) "HandySinger: Expressive singing voice morphing using personified hand-puppet Interface," *Proc. NIME2005*, Hamamatsu: 121–126.

豊田健一，片寄晴弘，河原英紀（2006）「STRAIGHTによる歌声モーフィングの初期的検討」『情報処理学会研究報告MUS』2006: 19, 59–64.

河原英紀，生駒太一，森勢将雅ほか（2007）「モーフィングに基づく歌唱デザインインタフェースの提案と初期検討」『情報処理学会論文誌』48, 3637–3648.

森勢将雅，河原英紀，片寄晴弘（2008）「STRAIGHTによるリアルタイム歌唱モーフィングシステムの実装」『情報処理学会研究報告MUS』2008: 50, 117–122.

河原英紀，森勢将雅，高橋徹ほか（2008）「実時間操作インタフェースへの応用を目的とした歌唱モーフィング操作パラメタの時系列への拡張について」『情報処理学会研究報告MUS』2008:127, 91–96.

Onishi, M., Takahashi, T. Irino, T. and Kawahara, H. (2008) "Vowel-based frequency alignment function design and recognition-based time alignment for automatic speech morphing," *Proc. SLT2008*, Goa India, 25–28.

西田沙織，大西壮昇，吉田有里ほか（2008）「STRAIGHTを用いた簡易モーフィングによる印象変化の評価について」『情報処理学会研究報告MUS』2008: 50, 43–48.

河原英紀，森勢将雅，高橋徹ほか（2009）「TANDEM-STRAIGHTおよび時変モーフィングのための研究用インタフェースの開発について」『電子情報通信学会技術報告』108: 465, 51–56.

Zen, H., Toda, T., Nakamura, M. and Tokuda, K. (2007) "Details of the Nitech HMMbased speech synthesis system for the Blizzard Challenge 2005," *IEICEJ Transactions*, E90-D, 325–333.

Toda, T., Black, A. W. and Tokuda K. (2007) "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *Trans. IEEE Audio, Speech and Language processing*, 15, 2222–2235.

Onishi, M., Takahashi, T., Morise, M., et.al. (2007) "Vowel-based voice conversion and its objective evaluation," *Proc. NCSP'08*, Goldcoast Australia, 275–278.

齋藤毅，後藤真孝，鵜木祐史，赤木正人（2008）「SingBySpeaking: 歌声知覚に重要な音響特徴を制御して話声を歌声に変換するシステム」『情報処理学会研究報告MUS』2008: 12, 25–32.

加藤和美，筧一彦（1996）「音声知覚における話者への適応性の検討」『日本音響学会誌』44, 180–186.

Schweinberger, S. R., Hauthal, N., Kaufmann, J., et. al. (2008) "Auditory adaptation in voice perception," *Current Biology*, 18, 684–688.

(Received Mar. 6, 2009, Accepted May 14, 2009)