Vol.2010-MUS-86 No.6
2010/7/28

歌声を見て触る: TANDEM-STRAIGHT と時変モーフィングが提供する基盤

河 原 英 紀[†1]　森 勢 将 雅[†2]

歌声情報処理研究の基盤として広く使われている STRAIGHT とモーフィングから，最新版の TANDEM-STRAIGHT および時変モーフィングのアルゴリズムと実装まで，その背景と併せて紹介する．STRAIGHT と TANDEM-STRAIGHT は，いずれも元の波形に含まれる位相情報を意図的に破壊しており，波形符号化方式とみなして評価すると SNR は −3 dB という劣悪な値になる．この破壊の代償として得られるものについての説明から，様々な応用のヒントをつかんで頂きたい．

Make singing voice tangible: TANDEM-STRAIGHT and temporally variable morphing as substrate

Hideki Kawahara[†1] and Masanori Morise[†2]

Algorithms and implementation details are introduced for latest TANDEM-STRAIGHT and temporally variable multi-aspect speech morphing, based on introduction of motivations behind the legacy-STRAIGHT and following developments. STRAIGHT and TANDEM-STRAIGHT intentionally destroy phase information in the original input speech. This destruction yields extremely poor SNR value (−3 dB) when they are evaluated as waveform coding methods. This article tries to illustrate views on prospective merits which this destruction provides in return. The authors introduced those views in the hope that readers of this article would be able to find interesting hints for their applications.

†1 和歌山大学
Wakayama University
†2 立命館大学
Ritsumeikan University

1. はじめに

音声分析変換合成法 STRAIGHT[1] は，聴覚や音声の研究の状況への欲求不満から生まれた[2]．STRAIGHT を応用した音声モーフィング[3]，様々な基本周波数をはじめとする音源情報の抽出法[4),5]，アルゴリズムを根本から見直した TANDEM-STRAIGHT[6]，TANDEM-STRAIGHT を応用した時変モーフィング[7] やツール[8] などの開発の背景には，この欲求不満を解消したいという想いが通奏低音となって流れている[⋆1]．

粘土の固まりを見つけたら手でこねて色々なものを作ることができる．紙と色鉛筆があれば絵を描いて視覚刺激を作ることができる．しかし音は見ることも触ることも変えることも簡単にはできない．

音声知覚研究の初期における sound spectrogram[10] と pattern playback[11] の出現は，「声を見て触って変える」手段を提供した（かに思えた）．しかし，Voder や channel vocoder[12] などと同様のアナログ技術に基づく方法で再現される劣悪な品質の音声[⋆2] は，高度に非線形な要素を含み多重の修復機構を内蔵する人間の音声知覚[13] [⋆3]を研究するためには，あまりにも粗かった．細密画を描くのに太いクレヨンしか使えないとしたら努力は報われるだろうか．

統計的モデルに基づくスペクトル推定法の発明[15] と，計算法としては等価な LPC の発明[16] により，分析合成の品質は大きく向上した[⋆4]．ただし向上したとはいえ，元の音声とは明らかに違う，いわゆる vocoder 声であることに変わりはない．それが，その頃の「分析合成系の品質には限界がある．高い品質には波形符号化が必須だ．」という通説につながる．

変な話だ．音声生成過程の構造は，vocoder に他ならない．同じ構造から出る音がなぜ，人間の肉声なら品質が高く，分析合成系では品質が悪いのか？異なった乱数から生成される二つの白色雑音は，同じ音に聴こえる．しかし，片方を信号と考え，二つの波形から SNR を求めると −3 dB という劣悪な値になる．波形符号化など波形の再現を評価関数とする

⋆1 23 年前にも同じようなことを書いていた[9] ことを思い出した．結局，考え方は進歩していないようだ．このときに作ったプログラムは，1989 年に「音声工房」という名前で NTT アドバンステクノロジから発売された．数百セット以上が出荷されたらしい．
⋆2 劣悪過ぎると，別の有用性が出て来る．VOCODER は，特有の劣化がエフェクターとして効果的に使用されているし，クロスシンセやしゃべる楽器にも VOCODER の原理が応用されている．
⋆3 人間の修復能力は現在の音声認識技術を凌駕している．音声 CAPTCHA への応用も検討されている[14]．
⋆4 有声音のパラメタを推定するのになぜ白色ガウス雑音により駆動されたモデルを用いなければならないのか，当時も今も気持が悪い．それ以外を仮定すると，面倒なことになることは理解しているが...

(SNR の改善を目的とする）やり方は，当面の戦術としては妥当であるとしても，何か本質を見逃しているのではないか？

「聴覚は位相を感じない (phase deaf)」という通説が本当であるなら，(短時間）パワースペクトルが知覚される音色を決めることになる．上で触れた白色雑音の問題も，パワースペクトルが（平均的に）同じだからとして説明できる．自己相関関数に基づく統計的な方法[15)16)] とも整合する．しかし，通説は真実ではない．反証は幾つも挙っている[17)18)]．何を信じたら良いのか．そのようなモヤモヤした気分の中で，STRAIGHT の元になるプログラムを書いてしまった．Vocoder なのに，音が良かった[⋆1]．そこで考えた[⋆2]．

2. 有声音を標本化機構とみなす

楽器音や歌声のようにほぼ周期的な信号は，滑らかで心地よく豊かなニュアンスを伴って聴こえる．しかし，その音を詳しく見ようとして，例えばスペクトログラムを表示すると，時間方向にも周波数方向にも，どこにも滑らかなものは見えない．この周期的駆動を，駆動対象の情報を標本化する操作だと解釈することで，背景にある滑らかなものを見ることができるようになる．これが STRAIGHT と TANDEM-STRAIGHT に共通するアイデアだ．

2.1 分析位置に依存しないパワースペクトル

短時間 Fourier 変換を用いて求められるパワースペクトルが時間方向で変動するのは，窓関数の周波数領域での表現の通過帯域内に複数の調波成分が含まれることによる．この問題には，通過帯域内に調波成分が一つしか含まれないように，十分に長い時間窓を用いるという自明な解がある．TANDEM[19)] は，この自明な解よりも短い窓を使って分析位置に依存しないパワースペクトルを求める方法である．後で説明するように，TANDEM には，その他にも自明な解には無い良さがある．

話を簡単にするために，窓関数の通過域は 2 個の調波成分を含む幅であり，サイドローブの影響は無視できるものとする．すると，一般性を失うこと無く，次のような信号 $x(t)$ を考えれば良く，さらに $k = 0$ と置くことができる．

$$x(t) = e^{jk\omega_0 t} + \alpha e^{j((k+1)\omega_0 t+\beta)} \tag{1}$$

ここで α, β は，適当な実数であり，$\omega_0 = 2\pi f_0 = 2\pi/T_0$ は，基本周波数 f_0 に対応する角周波数を表す．窓関数の Fourier 変換を $W(\omega)$ とすると，スペクトル[⋆3] $P(\omega, t)$ は，次のようになる．なお $k = 0$ と置いた．

$$P(\omega,t) = |W(\omega)|^2 + \alpha^2 |W(\omega-\omega_0)|^2 + 2W(\omega)W(\omega-\omega_0)\cos(\omega_0 t+\beta)\ , \tag{2}$$

第三項が時間変動する成分を表す．この成分は周期が T_0 の余弦波なので，以下のように $T_0/2$ の時間を隔てた位置で求めたスペクトログラムとの平均を計算することで，消去することができる．

$$P_T(\omega,t) = \frac{1}{2}\left[P\left(\omega, t-\frac{T_0}{4}\right) + P\left(\omega, t+\frac{T_0}{4}\right)\right]\ . \tag{3}$$

このように定義される $P_T(\omega, t)$ を TANDEM スペクトルと呼ぶことにする．

2.1.1 窓関数の選択

窓関数の周波数領域での表現にはサイドローブがある．そのため，実際には $P_T(\omega, t)$ も時間的に変動する．細かな議論を省いて結論を言えば[⋆4]，窓長が $2.5T_0$ の Blackman 窓が最も実用的ということになる．

音声分析では対数スペクトルの性質が問題となる．ここでは，対数スペクトルの変動を表す指標 η_{dBt} を，以下のように定義して雑音の影響を評価することとする．

$$\eta_{dBt} = \sqrt{\left\langle \frac{1}{2\pi T_0}\int\int_0^{T_0} \left|L(\omega,t) - \overline{L(\omega)}\right|^2 dt\ d\omega \right\rangle} \tag{4}$$

なお，式中の $\overline{L(\omega)}$ は，以下で定義される．

$$\overline{L(\omega)} = \left\langle \frac{1}{T_0}\int_0^{T_0} L(\omega,t)\ dt \right\rangle,\ L(\omega,t) = 10\log_{10} P(\omega,t), \tag{5}$$

ここで，$\langle X\rangle$ は，X の期待値（実際には乱数を用いたシミュレーションにより求めた平均値）を表す．

図 1 に，結果の一例を示す．左側が自明な解を用いた場合の結果，右側が TANDEM の結果である．入力信号には周期を T_0 とするパルス列を用い，設定した SNR となるように白色ガウス雑音を加えた．横軸は，2 次のモーメントから求めて基本周期で正規化した持続時間 σ_t，縦軸は対数スペクトルの変動量 η_{dBt} であり，設定する SNR の値として 30 dB を用いた．図では比較のために，Blackman 窓に加え，Hanning 窓，Kaiser 窓（$\beta = 9$）と[21)]，

⋆1 現在の STRAIGHT と比べると，ひどい品質でしかない．
⋆2 勧められるやり方ではないかも知れないが，動くプログラムを書いてしまってから考えている．STRAIGHT 関連の論文は，後付けの説明が多い．説明を間違えていることもある．Matlab code の方を信用して欲しい．
⋆3 正しくはスペクトログラムだが，混乱しない限り，以下ではスペクトルと呼んでおくことにする．
⋆4 細かな議論の一部は，文献 20) にある．

Vol.2010-MUS-86 No.6
2010/7/28

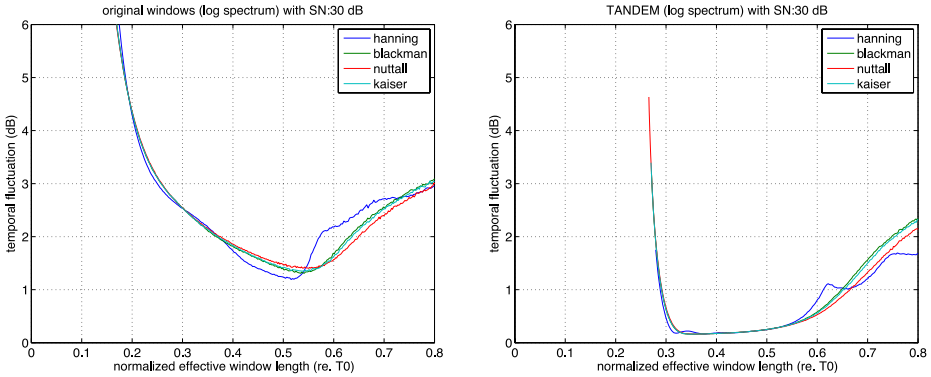

図 1 Temporal variation of logarithmic power spectra under different SNR. (left) original time windows. (right) TANDEM windows. The SNR is 30 dB

Nuttall 窓[22] の結果を併せて示している．Blackman 窓を用いた場合の最良の条件 $2.5T_0$ に対応する $\sigma_t = 0.388$ 付近では，TANDEM を用いた場合の対数スペクトルの変動量は，自明な解を用いた場合の約 1/10 になっている．これは，Cepstrum を計算する際に非常に都合の良い性質である．また，図から分かるように，分析対象とする入力信号の周期と設計に用いた T_0 との誤差が約 12%以内であれば，時間変動は実質的に増加しない．この許容範囲は，約 4 半音（400 cent）に相当する．TANDEM は，かなりいいかげんな使い方でもメリットがあることになる．さらに，T_0 が分からない最悪の場合でも，$N = 2$ の Welch 法[23] としてのメリットは残る．

2.2 周波数方向の変動の除去

こうして求められた TANDEM スペクトルには，周波数軸上で周期を f_0 とする振動が重なっている．この成分は，空間周波数 $1/f_0 = T_0$ に零を持つような平滑化関数をスペクトルに畳込むことで除去できる．そのような関数の最も簡単なものは，幅が f_0 の矩形である．矩形の平滑化関数を用いて変動成分を消去した平滑化スペクトル $P_S(\omega, t)$ は，次式で求められる．

$$P_S(\omega, t) = \frac{1}{\omega_0} \int_{-\frac{\omega_0}{2}}^{\frac{\omega_0}{2}} P_T(\omega - \lambda) d\lambda \tag{6}$$

しかし，こうして求めた $P_S(\omega, t)$ は，窓関数の周波数応答による平滑化と，矩形の平滑化関数による平滑化の影響が二重に含まれていることになる．

ある応答を持つシステムを周期的なパルスで駆動することは，周波数領域で応答に対応するスペクトルを周期的に標本化することでもある．上の問題は，理想的ではない anti-aliasing filter を用いた A/D，D/A 変換系での元の波形の復元の問題とみなすことができる．consistent sampling[24] は，このような状況で，D/A 変換後に再度標本化された値を元の信号を標本化した値と一致させる方法を与える．

具体的には，係数 q_k で決まる周波数領域のデジタルフィルタを用いて，平滑化スペクトル $P_S(\omega, t)$ を処理することにより，補償されたスペクトル $P_{ST}(\omega, t)$ が求められる．

$$P_{ST}(\omega, t) = \sum_{k=-\infty}^{\infty} q_k P_S(\omega - k\omega_0, t) \tag{7}$$

係数 q_k は，平滑化関数 $h(\omega)$ と，窓関数の周波数領域での表現 $W(\omega)$ から，次式により求められる．

$$Q(z) = \frac{1}{R(z)} = \frac{1}{\sum_{k=-\infty}^{\infty} r_k z^{-k}} = \sum_{k=-\infty}^{\infty} q_k z^{-k} \tag{8}$$

$$r_k = \int_{-\infty}^{\infty} h(\omega - k\omega_0) \left| W(-\omega) \right|^2 d\omega,$$

ここで $h(\omega)$ を前述の矩形関数とし，窓関数を長さが $2.5T_0$ の Blackman 窓とすると，$|r_k|$ および $|q_k|$ の値は，$|k|$ の増加とともに急速に減少することが分かる．したがって，実用上は $|k| < 2$ の係数を考慮するだけで良い．

2.2.1 正値性の保証と実装

この処理には，副作用がある．処理で用いる q_1 と q_{-1} は負の値となるため，求めた $P_{ST}(\omega, t)$ が常に正の値をとることを保証できなくなる．その結果，$P_{ST}(\omega, t)$ に正以外の値が含まれることになると合成のための最小位相応答[⋆1] を計算することができないという問題が生ずる．

TANDEM スペクトルでは，調波成分のパワーと比較すると，信号の周期性によるスペクトルの周期的な変動は遥かに小さい．そこで，$|x| \ll 1$ の場合に $\log(1 + x) \approx x$ である

⋆1 零位相で合成した音の品質は，明らかに悪い．しかし，零位相と最小位相応答による合成音声の品質の差を系統的に評価した報告を，残念ながら見つけることができなかった．

ことを利用して，処理を対数スペクトルの上で行うこととした．処理結果を指数関数で変換して $P_{ST}(\omega,t)$ とすることで，正値性を保証することができる．これらをまとめると，式 6 と式 7 の代わりに，以下を計算することになる．

$$L_S(\omega,t)=\frac{1}{\omega_0}\int_{-\frac{\omega_0}{2}}^{\frac{\omega_0}{2}}\log\left(P_T(\omega-\lambda)\right)d\lambda \tag{9}$$

$$P_{ST}(\omega,t)=\exp\left(q_0L_S(\omega)+q_1(L_S(\omega-\omega_0,t)+L_S(\omega+\omega_0,t))\right) \tag{10}$$

ここで $q_1=q_{-1}$ を利用した．なお，実際には式 9 の平滑化の処理と，式 10 の補償の処理を一括して，cepstrum llifter として実装している．このようにして求められる $P_{ST}(\omega,t)$ を，STRAIGHT スペクトルと呼ぶことにする[⋆1]．

2.3 混 合 音 源

音声の再現や変換のためには，この STRAIGHT スペクトルから構成されるフィルタを駆動する信号が必要となる．TANDEM-STRAIGHT では，周期的パルスと広帯域の有色雑音からなる混合音源が用いられている．パルスの繰返し周期は，抽出された基本周波数[20]を用い，雑音の特性は，周期／非周期の境界周波数と境界での遷移の傾斜の二つのパラメタを用いて定められる[25]．なお，パルスの繰返し周期は群遅延を操作することにより，標本化周期よりも細かな時間分解能で設定される．これらの詳しい説明は，それぞれの参考文献に譲る[⋆2]．音源については，さらに，強い印象を与えるだみ声やシャウトなどの自由な操作を可能にするための検討が進められている．

3. もう一つの背景

このように音声を分析して得られるフィルタ情報と音源情報とを組み合わせて，元の音声と同等の自然性を有する音声を再合成することのできる STRAIGHT は，高い品質での音声のモーフィングを可能にした．モーフィングを利用すると，知覚的印象の異なる二つの事例を用意するだけで，知覚的属性と物理特性との対応関係に関する事前知識が無い場合にでも，二つの事例を結ぶ刺激連続体を作ることができる．このような刺激連続体は，音声の知覚研究の手段[26]としてだけではなく，演奏表現の新しい操作手段[27]やコンテンツ制作の素材としても応用されている[⋆3]．

事例間を結ぶ経路は，自由に決めて良い．パラメタの組合せを適切に選択することにより，「歌い回し」と「声質」を独立に操作することができる[30]．これを応用すると，例えば「この歌手の歌い回しと，あの歌手の声質をこんな風に混ぜて...」のように事例を直接参照したコンテンツの加工が可能になる．このような操作感を先行してインタラクティブに提供したインタフェース v.morish[31] は，高い関心を集めた[⋆4]．時変モーフィング[⋆5]は，このインタフェースをポストプロダクションのオートメーションなどで利用するシステムや，ライブでの使用が可能なリアルタイムシステムとして実現するために必要な基盤を提供する．

4. 時変モーフィング

事例を参照した自由な加工では，加工操作が外挿の領域に入っても破綻しないことと，加工の内容を（時間的に）局所的に変えられることが必要になる．モーフィングを線形補間として実装する方法[3] は，いずれの要求を満たすこともできない．

4.1 外挿で破綻しないモーフィング

「外挿で破綻しない」という条件を満たすために，まず，以下に示すように，導関数の対数の線形補間の指数関数による変換としてモーフィングの定義を変更する．

$$T_{Am}(x_A)=\int_0^{x_A}\exp\left(\log\left(\frac{dT_{Am}(\lambda)}{d\lambda}\right)\right)d\lambda=\int_0^{x_A}\left(\frac{dT_{AB}(\lambda)}{d\lambda}\right)^{r_{AB}}d\lambda\ , \tag{11}$$

込み入った議論になるので，以下のように整理した表記法を用いている．モーフィングの対象となる事例を A, B とする．時間軸や周波数軸等，それぞれの事例のフィルタパラメタや音源パラメタの定義されている座標を表す変数を，事例の添字を付けて x_A, x_B のように表す．事例 A の座標を事例 B の座標に変換する変換を $T_{BA}(x_A)$ のように変換後と変換前の添字をこの順序で付けて表す．添字 m を，モーフィングされた結果を表す座標や変換の添字として用いる．事例 B を基準としたモーフィングで事例 B を事例 A に完全に移す変換の場合のモーフィング率 r_{BA} を 1，事例 A に移す場合の率を 0 と定義する．この場合，逆

⋆1 実際には，q_k の打切りや対数の Taylor 展開の高次項の影響，聴覚末梢系での情報表現と対数スペクトルの違いなどにより，q_k の $|k|<2$ の項をそのまま用いることはできない．補償用係数の適切な設定については，別の機会に報告する．

⋆2 いろいろなところに情報が分散しているのには著者自身も困っているので，網羅的な資料を執筆している．今年中に掲載されることになるはずです...

⋆3 例えば，未来館の企画展[28]では，声優により演じられた「喜び」「哀しみ」「怒り」の感情音声をインタラクティブにモーフィングする作品が展示された．ここでは三つの事例から 139 種類のモーフィング音声が予め作成され，Flash ムービーに埋め込まれた．デモのコピーへのリンクを 29) にあげる．

⋆4 v.morish の動作しているデモをニコニコ動画で見ることができる[32]．

⋆5 英文では時変多属性モーフィングとしていたが，ここでは誤解の無い限り，略称を用いる．

方向で見たモーフィング率として定義される r_{AB} の値は，それぞれ 0, 1 となる．なお，恒等写像の導関数の対数が 0 となることが式の簡単化に利用されている．

4.2 リアルタイムシステム

v.morish のようにリアルタイムでモーフィング率を操作する場合には，例えば $r_{BA}(t_s)$ のようにモーフィング率が現実の時間軸 t_s 上の関数となる．この場合には，次式に示すように，t_s を逐次更新しながら，現在の時刻が A の事例の上でのどの時刻に対応するかを表す変換関数 $T_sA(t_s)$ と，同様に B の時刻への変換関数 $T_sB(t_s)$ を逐次更新する．

$$t_s = \int_0^{t_s} d\lambda, \tag{12}$$

$$T_{sA}(t_s) = \int_0^{t_s} \left(\frac{dT_{AB}(T_{sA}(\lambda))}{d\lambda} \right)^{-r_{AB}^{(t)}(\lambda)} d\lambda, \tag{13}$$

$$T_{sB}(t_s) = \int_0^{t_s} \left(\frac{dT_{BA}(T_{sB}(\lambda))}{d\lambda} \right)^{(r_{AB}^{(t)}(\lambda)-1)} d\lambda, \tag{14}$$

これらを用いて t_s に対応する事例上の時刻 $T_{sA}(t_s)$, $T_{sB}(t_s)$ を使って，素材とするパラメタを読み出し，混合する．

$$\Theta_m(t_s) = (1 - \vec{r}_{AB}(t_s))\Theta_A(T_{sA}(t_s)) + \vec{r}_{AB}(t_s)\Theta_B(T_{sB}(t_s)). \tag{15}$$

ここで $\Theta(t)$ は，時刻 t でのパラメタの組を表す．モーフィング率は，パラメタおよび軸のそれぞれに別の値を設定して構わないので，$\vec{r}(t)$ のように，要素が時間の関数であるベクトル量として表す．

4.3 非リアルタイムオフラインシステム

ポストプロダクションのように，非リアルタイムでのモーフィングでは，その上でモーフィング率を定義するための参照用の時間軸 t_r が必要になる．この t_r 軸上で設定された時間軸のモーフィング率 $r_{rAB}^{(t)}(t_r)$（これも時間の関数）を用いて，参照用の時間軸上の時刻を事例の時間軸上の時刻に変換した $T_{rA}(t_r)$, $T_{rB}(t_r)$ によりパラメタ等を読込み，混合した結果を $T_{rs}(t_r)$ により，現実の時間 t_s のパラメタとする．詳細は文献 7) にゆずる．

5. 実装と GUI

TANDEM-STRAIGHT と時変モーフィングは，科学技術計算用の環境である Matlab を用いて開発されており，ここまでで紹介したアルゴリズムが関数群として提供されている．これらの関数は，速度よりも可読性と可用性を重視したコードで実装されており，プラットフォームに依存しない．様々なシステムへの応用[⋆1]では，基盤となるそれらの関数群を構成する関数への呼び出しを組み合わせて用いることが想定されている[⋆2]．

6. おわりに

TANDEM-STRAIGHT や時変モーフィングは，基盤／基層（substrate あるいは substratum）である．ぜひ，その基盤の上に，新しい発想で様々な応用システムや斬新なインタフェースを開発して歌声の愉しみをより豊かなものにして頂きたい．これまでの説明で分かるように，特定の応用を想定した場合には，この基盤には過剰品質となっている部分が多い．また，Matlab コードには，特許を回避したりオリジナリティーを主張するためにそれほど有効ではない捻りを加えてある部分もある．使用にあたっては，ぜひコードを批判的に読んで理解してそれらを仕分けて欲しい．信号処理の基盤は，音声認識や音声合成システムと違い，個人で全体を完全に把握することが容易である．この発表の前の発表[34] のように，最初から自分のアイデアを加えて作り直すことも試みて欲しい．

なお，現在の基盤では，まだ最初に触れた欲求不満を解消するには力不足である．SNR を定義することが無意味なほど悪い SNR ではあるけれども知覚的には元の声と区別できない品質の音声を，知覚的に意味のあるパラメタで記述し操作できる基盤の実現を目標に，さらに研究を進めて行きたい．

謝辞 STRAIGHT を発端として TANDEM-STRAIGHT と時変モーフィングおよび応用技術へと続く一連の研究は，様々な支援を受けて進められて来た．本資料で紹介した最近の研究は，主に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 プロジェクトと，科学研究費 基盤 (A)19200017 の支援によるものである．

参考文献

1) 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).
2) 河原英紀：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

⋆1 本研究会で発表されるものを含め，STRAIGHT は多くのシステムで用いられている．それらのリストは，STRAIGHT の紹介ページ[33] からのリンクを参照願いたい．
⋆2 時変モーフィングでは設定すべきパラメタの数が多く手続きも込み入っているため，我慢できずに，GUI を備えたツール[8] を開発してしまった．このツールは，基盤ではない．応用システムの実装の例と考えて欲しい．

STRAIGHT の背景と展開 –，日本音響学会誌， Vol.63, No.8, pp.442–449 (2007).
3) Kawahara, H. and Matsui, H.: Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation,, *ICASSP'2003*, Vol.I, pp.256–259 (2003).
4) 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, *EUROSPEECH'99*, Vol.6, pp.2781–2784 (1999).
5) Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T. and Irino, T.: Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT, *Interspeech'2005*, pp.537–540 (2005).
6) Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. and Banno, H.: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation, *ICASSP'2008*, pp.3933–3936 (2008).
7) Kawahara, H., Nisimura, R., Irino, T., Morise, M., Takahashi, T. and Banno, B.: Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown, *ICASSP2009*, pp.3905–3908 (2009).
8) Kawahara, H., Takahashi, T., Morise, M. and Banno, H.: Development of exploratory research tools based on TANDEM-STRAIGHT, *APSIPA'2009*, pp.111–120 (2009).
9) 河原英紀：音声知覚過程研究支援環境のユーザインタフェース，聴覚研究会資料，Vol.H-87-21 (1987).
10) Koenig, W., Dunn, H.K. and Lacy, L.Y.: The sound spectrograph, *J. Acoust. Soc. Am.*, Vol.18, No.1, pp.19–49 (1946).
11) Liberman, A.M., Delattre, P.C. and Cooper, F.S.: The rôle of selected stimulus-variables in the perception of the unvoiced stop consonants, *American Journal of Psychology*, Vol.65, pp.497–516 (1952).
12) Dudley, H.: Remaking Speech, *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177 (1939).
13) 柏野牧夫：音韻修復–消えた音声を修復する脳–，日本音響学会誌， Vol.61, No.5, pp. 263–268 (2005).
14) 西本卓也，松村 瞳，渡辺隆行：音声 CAPTCHA における了解度と心的負荷の検討，音響学会春季研究発表会， No.3-4-3, p.121 (2010).
15) 板倉文忠：統計的手法による音声スペクトル密度とフォルマント周波数の推定，電子情報通信学会論文誌 A， Vol.53-A, No.1, pp.35–42 (1970).
16) Atal, B.S. and Hanauer, S.L.: Speech analysis and synthesis by linear prediction of the speech wave, *J. Acoust. Soc. Am.*, Vol.50, No.2B, pp.637–655 (1971).
17) Plomp, R. and Steeneken, H. J.M.: Effect of Phase on the Timbre of Complex Tones, *J. Acoust. Soc. Am.*, Vol.46, No.2B, pp.409–421 (1969).
18) Patterson, R.D.: The sound of a sinusoid: Spectral models, *J. Acoust. Soc. Am.*, Vol.96, No.3, pp.1409–1418 (1994).
19) 森勢将雅， 高橋徹，河原英紀，入野俊夫：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，電子情報通信学会論文誌 D， Vol.J 90-D, No.12, pp.3265–3267 (2007).
20) 河原英紀，和田芳佳，森勢将雅，西村竜一，入野俊夫：音源構造抽出法の初期推定値のバイアス除去と高速化について，聴覚研究会資料 (2010). （2010.7.17 発表予定）.
21) Harris, F.J.: On the use of windows for harmonic analysis with the discrete Fourier transform, *Proceedings of the IEEE*, Vol.66, No.1, pp.51–83 (1978).
22) Nuttall, A.H.: Some windows with very good sidelobe behavior, *IEEE Trans. Audio Speech and Signal Processing*, Vol.29, No.1, pp.84–91 (1981).
23) Welch, P.: 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*, Vol.15, No.2, pp.70 – 73 (1967).
24) Unser, M.: Sampling–50 Years After Shannon, *Proceedings of the IEEE*, Vol. 88, No.4, pp.569–587 (2000).
25) 河原英紀，森勢将雅，高橋 徹，坂野秀樹，西村竜一，入野俊夫：高品質分析合成のための有声音の非周期成分の表現と推定について，聴覚研究会資料 H-2010-44，Vol.40, No.3, pp.231–236 (2010).
26) Schweinberger, S.R., Casper, C., Hauthal, N., Kaufmann, J.M., Kawahara, H., Kloth, N., Robertson, D.M., Simpson, A.P. and Zaeske, R.: Auditory Adaptation in Voice Perception, *Current Biology*, Vol.18, No.9, pp.684–688 (2008).
27) Yonezawa, T., Suzuki, N., Abe, S., Mase, K. and Kogure, K.: Perceptual continuity and naturalness of expressive strength in singing voices based on speech morphing, *EURASIP Journal on Audio, Speech, and Music Processing*, No.3 (2007).
28) :「恋愛物語展」 ーどうして一人ではいられないの？(2005.4.15〜2005.8.15）.
29) : http://www.wakayama-u.ac.jp/%7ekawahara/Miraikandemo/straightMorph.swf.
30) 河原英紀，生駒太一，森勢将雅，高橋 徹，豊田健一，片寄晴弘：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， Vol. 48, No.12, pp.3637–3648 (2007).
31) Morise, M., Onishi, M., Kawahara, H. and Katayose, H.: v.morish'09: A morphing-based singing design interface for vocal melodies, *Lecture Note in Computer Science*, No.LNCS 5709, pp.185–190 (2009).
32) : http://www.nicovideo.jp/watch/sm4747100.
33) : http://www.wakayama-u.ac.jp/%7ekawahara/STRAIGHTadv/index_j.html.
34) 森勢将雅，中野皓太，西浦敬信：実時間歌唱力補正に基づく新たなカラオケエンタテインメントの創出，音楽情報科学研究会， No.MUS86-6 (2010).