音楽情報科学 24−13
(1998. 2. 20)

音声分析変換合成法 STRAIGHT における音源情報の精密化について

河原英紀[1,2]、 片寄晴弘[1,3]

[1]和歌山大学/CREST
[2]ATR 人間情報通信研究所 [3]イメージ情報研究所
[1]〒 640-8510 和歌山市栄谷 930
kawahara@sys.wakayama-u.ac.jp

あらまし 高品質な音声の分析変換合成方法として提案された STRAIGHT (Speech/sound Transformation and Representation using Adaptive Interpolation of weiGHTed spectrogram) における幾つかの品質劣化要因を解析し対策を検討したので報告する。一つは有声／無声判定の処理を不要とし、すべての音源情報を連続量として表したことである。もう一つは、ピッチ周期の間で複数回の励起を有する場合に生ずるスペクトル上の二次的構造を除去したことである。これらの改良により、処理の頑健性が向上し以前の変換音声に認められた品質上の問題が解消された。

キーワード 駆動信号、周期性、音声分析、変換、合成

Refinement of source representations in a speech analysis-modification-synthesis method, STRAIGHT

Hideki Kawahara[1,2] and Haruhiro Katayose[1,3]

[1]Wakayama University/CREST, [2]ATR Human Information Processing Research Laboratories,
[3]Laboratory of Image Information Science and Technology
[1]930 Sakaedani, Wakayama, Wakayama 640-8510
kawahara@sys.wakayama-u.ac.jp

Abstract Refinement of a high-quality speech analysis, modification and synthesis method, STRAIGHT (Speech/sound Transformation and Representation using Adaptive Interpolation of weiGHTed spectrogram), in terms of excitation source representations was conducted to overcome several defects in reproduction quality. One is a multi-band mixed excitation source with continuous weighting factor was introduced to eliminate a hard voiced/unvoiced decision. The other is a suppression mechanism to reduce a second order interference caused by multiple excitation within one pitch period. These modification were found to make STRAIGHT robust and to reduce quality deficiencies.

key words excitation, periodicity, analyisis, modification, synthesis

1 はじめに

STRAIGHT は、広範囲の音声パラメタの操作と高品質な音声の分析合成とを両立させた、channel VOCODER[2] の一種である [4, 7, 5]。STRAIGHT は、有声音のようなほぼ周期的な駆動による時間周波数表現への干渉を、基本周期の情報を用いて組織的に除去する手続 (STRAIGHT-core)[4, 10] と、「基本波らしさ」という尺度を用いて基本波の瞬時周波数として基本周波数を抽出する手続（TEMPO: Time domain Excitation extractor using Minimum Perturbation Operator）[9] と、合成用駆動インパルスの群遅延の変形を用いた品質の操作手続（SPIKES: Synthetic Phase Impulse for Keeping Equivalent Sound）[4, 12] を併用することで、VOCODER の品質上の問題点を解消している。また、周波数領域ならびに時間領域での処理を合成の際に加えることにより、分析の段階で加わる過剰平滑化を補正する方法を明らかにした [11, 13]。これらの改良を取り入れた STRAIGHT による女声の分析合成音声は、原音声に匹敵する高い自然性を有することが報告されている [13]。しかし、男声の分析合成では、しばしばバズのような音が聞こえたり、基本周波数を変換した場合に音が濁ったり、子音部が有声摩擦音のような響を持つ場合があり、自然性を損なっていた [1]。また、標本化周波数が低い（例えば 8kHz）場合の品質は、特に際立ったものではなかった [8]。

これらに加え、STRAIGHT のこれまでの実装では、有声摩擦音に相当するモデルを持たない等、本格的な音声知覚研究用のツールとして用いようとする場合 [14] に不備な点があった。本報告では、これらの問題点を解消することを目的として導入した幾つかの改良について説明する。

2 STRAIGHT における品質劣化

ここでは、STRAIGHT を用いた分析合成音声で認められた品質上の主要な問題点について、具体例に基づいて紹介する。

2.1 不十分な音源モデル

品質上問題を生ずる音声があったとき、時間軸を 10 倍程度に拡大して分析合成音声を作成すると、どのような状況で問題が生じているかを特定することが容易になる。予備的な検討の結果、バス音の発生は、主に有声音終了部周辺に集中していることが認められた。このような部分では、基本波領域に周期性が認められても、高い周波数領域では周期成分はほとんど存在していない。また、'right' の [t] のように有声音に後続する無声破裂音が有声摩擦音のように変化する例も認められた。この場合も、基本周波数領域における周期性（基本波らしさ）の残留と、高い周波数領域での非周期性という同様な状況で問題が生ずることが観察された。

女性の音声の場合には、基本周波数の存在しない低い周波数領域における空調雑音やハムの混入により、それらの周期成分を音声の基本周波数として誤認してしまう事例も認められた。

2.2 スペクトルの 2 次構造

男性の音声では、特に声に張りのある場合、基本周波数（F0）を変換した場合に音が濁る問題にしばしば遭遇した。パラメタ変換を行わない場合にはそのような品質上の問題は生じないため、STRAIGHT-core で求められた時間周波数表現に周期性に起因する干渉が何らかの形で残留していることが示唆される。実際、問題を有する男声の場合には、平滑化処理を経たスペクトログラムにおいても、F0 の動きに相似なスペクトル上の横縞状構造が認められる。

2.3 分析例

以上のような問題を有する音声の一例について、音源情報の分析例を図 1に示す。音声は、米国人男性の発声した "right" である。図 1では、上から順に、音声波形、パワー、基本周波数、基本波らしさ、基本波のパワー、帯域毎の基本波らしさのマップを示している。有声音の声門の閉止が存在しなくなったと考えられる 500ms あるいは 530ms 以降になっても、声帯は振動しているようであり、基本周波数は連続的に推移し、「基本波らしさ」もある程度の値を保っている。そのため、適応的な時間窓のサイズがこの低域成分の周波数によって決まると、高い周波数に主要な成分を有する破裂音の場合であっても、必要以上に時間分解能が低下し、必要以上に周波数分解能が向上するという問題が生ずる。

同じ音声資料を、TEMPO による F0 情報で決まる適応的な窓長の Gauss 窓を用いて分析した結果を図 2に示す。上で指摘したように 600ms 付近にある [t] の破裂の部分では、70Hz 程度の周期的な成分に対応する長い時間窓による分析が行われており、時間分解能が低下していることが分かる。

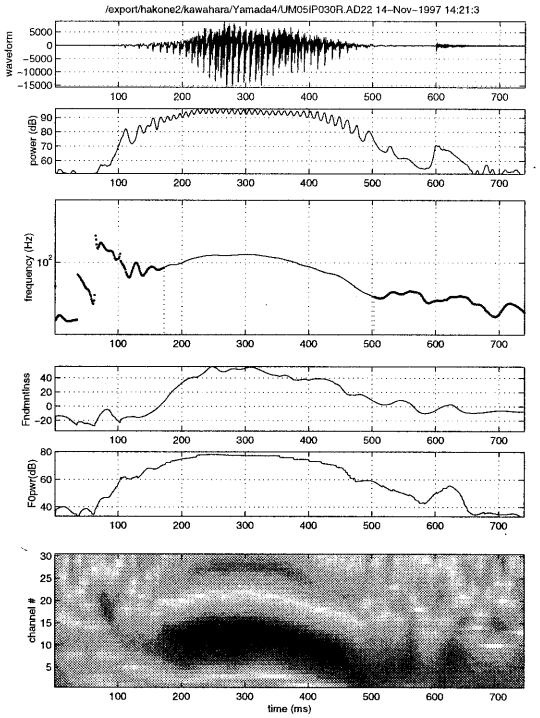

図 1: 男性の発声した “right” の音源情報の分析例

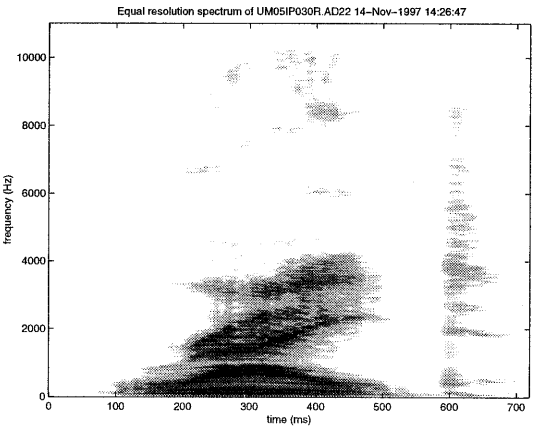

図 2: 男性の発声した “right” の F0 適応 SFFT 分析結果。F0 による干渉（ただし、印刷上では不鮮明）の他に、2 本の調波を単位とする干渉が認められることに注意。

100ms から 500ms に至る有声音の部分では、F0 による干渉が明瞭に見える（ただし、印刷上では不鮮明）のと同時に、F0 の 2 倍の間隔での干渉構造も見える。この干渉は、2 本の調波が組となった形で現われており、声帯音源の周期性に関連したものであることが分かる。有声音の終端部分にあたる 500ms 付近では、低い周波数領域では声帯は周期的に振動しているものの、高い周波数領域での周期性は、ほとんど認められない。

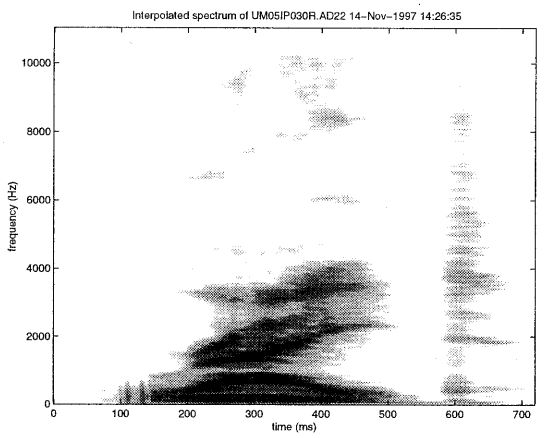

図 3: 男性の発声した “right” の STRAIGHT-core による処理結果。F0 による干渉は除去されているが、2 次的な構造は明瞭に残っている。

このスペクトログラムを STRAIGHT-core により平滑化した結果を図 3に示す。F0 による干渉構造は消えているが、2 本の調波を単位とする干渉は明瞭に残っている。以下では、F0 による直接的な干渉を除去した後に残るこのような構造のことを、便宜的に 2 次構造と呼んでおくこととする。このように 2 次構造が残留していると、元の音声の基本周期の半分の時刻の周辺に副次的な応答成分が発生することになる。F0 を変換して合成音声を作成した場合には、元の F0 に関連したこの副次的な応答と変換した F0 に関連した応答とが混在することになる。また、これまでの報告で提案してきた最適平滑化関数や時間領域処理による補正は、これらの 2 次構造を強調する方向に働く。

2.4 スペクトル 2 次構造

スペクトログラム上では、時間方向の相関が視覚的に利用できるため、2 次構造は明瞭に観察される。しかし、ある時刻のスペクトルだけを取り出した場合には、cepstrum を用いても、周波数軸上で線形予

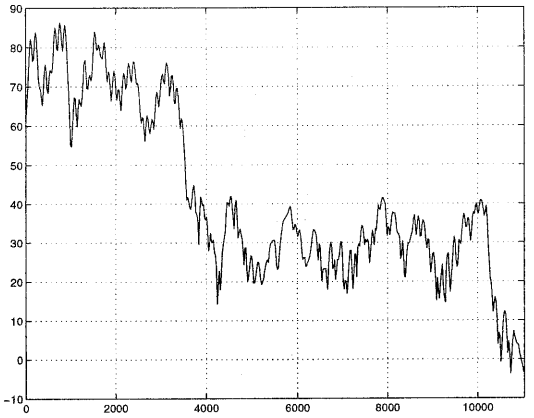

図 4: 男性の発声した “right” の 330ms の時点におけるパワースペクトル。

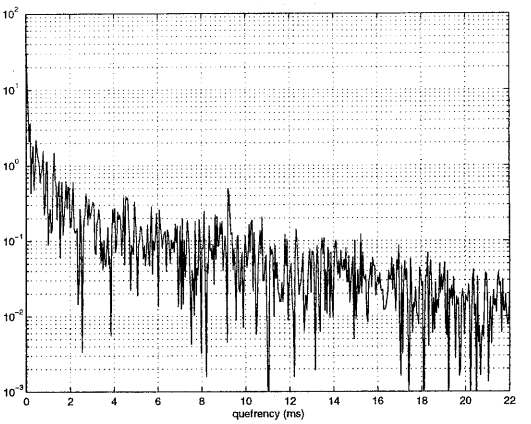

図 6: 男性の発声した “right” の 330ms の時点における cepstrum。

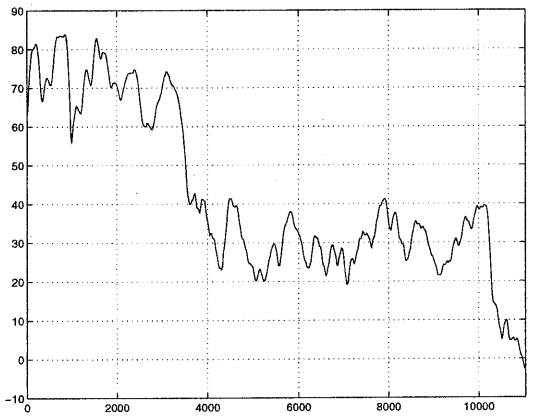

図 5: 男性の発声した “right” の STRAIGHT-core による処理結果の 330ms の時点における平滑化パワースペクトル。

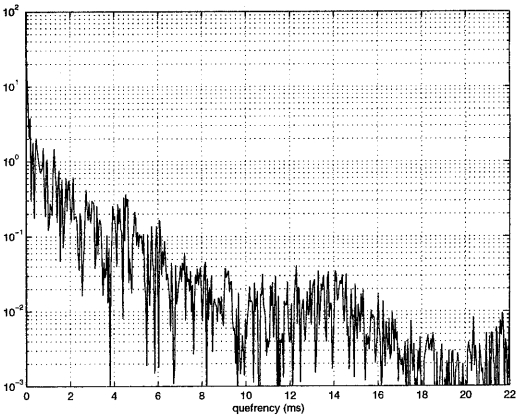

図 7: 男性の発声した “right” の STRAIGHT-core による処理結果の 330ms の時点における平滑化パワースペクトルから求めた cepstrum。

測分析を行っても検出は困難である。具体例を同一の音声の 330ms 位置におけるスペクトルを例として説明する。

図 4は、男性の発声した “right” の 330ms の時点におけるパワースペクトルである。F0 による干渉構造が明瞭に観測される。図 5は、男性の発声した “right” の STRAIGHT-core による処理結果の 330ms の時点における平滑化パワースペクトルである。STRAIGHT-core の平滑化により、F0 による干渉構造はほぼ完全に除去されている。しかし、スペクトログラムの所で説明したように、ピッチ周期以下の構造による干渉とおぼしきものの影響が見える。ただし、スペクトログラムを見ていなければ、それがピッチ周期に関連したものであることを理解することは困難である。つまり、この図の解析のためには、F0 についての事前知識が必要なのである。

図 6は、男性の発声した “right” の 330ms の時点における cepstrum である。F0 に対応する 9.2ms の quefrency でのピークが明瞭に観測される。図 7は、

男性の発声した "right" の STRAIGHT-core による処理結果の 330ms の時点における平滑化パワースペクトルから求めた cepstrum である。F0 によるピークは、完全に除去されている。cepstrum の形状は、cardinal B-spline 平滑化演算の特性を反映して、n/F0 の quefrency の周囲で低下している。

しかし、この cepstrum からは、平滑化スペクトログラムで認められていたピッチ周期内での駆動による干渉の構造から想定されるピークは、1/2F0 に相当する 4.6ms の位置にはほとんど見えない。強いて言えば、4.5ms を中心とした成分が全体としてやや大きくなっているように見えるだけである。このようなピークの有無を客観的に判定するのは、かなり困難であろう。

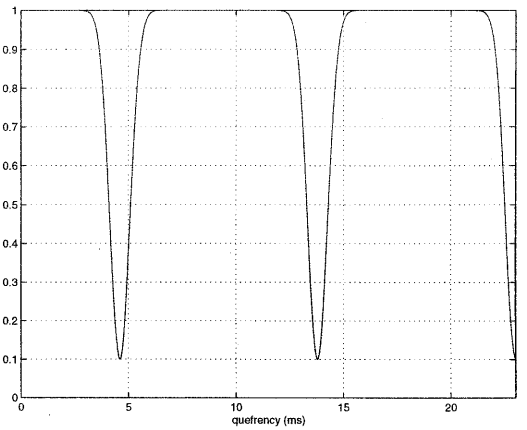

図 8: 基本周期 9.2ms に対応する 2 次構造抑圧用の cepstrum lifter。

3 スペクトル 2 次構造の除去

以上、見てきたように、2 次構造を単一の時点でのスペクトルだけから安定して検出することは難しい。一方、主要な駆動による応答が減衰しているピッチ周期の中間においては、付加的な音の検出感度が高くなるため [6]、2 次構造による副次的な応答による品質への影響は、平均的な S/N から想定されるものよりも高くなる可能性がある。

ここでは、予備的な検討の結果見いだされた 2 次構造が 2 つの調波が組になって生ずるものであるという先見的な知識を利用し、基本周期に相当する quefrency の 1/2 の奇数倍の quefrency において滑らかな極小値を有するような cepstrum lifter によって 2 次構造を抑圧することを試みる。具体的には $(1+\cos x)^{\beta}$ が β の増加とともに原点付近で Gauss 関数に漸近することを利用して次式を用いた。

$$w_{cep}(q) = 1 - \alpha(1 - 0.5(1 + \cos\frac{2\pi q}{\tau_0}))^{\beta} \quad (1)$$

ここで、α は 2 次構造の抑圧の度合いを決める係数、q は quefrency、τ_0 は、基本周期、β は、抑圧領域の幅を決める係数である。具体的な加重の例を図 8に示す。ここで、$\alpha = 0.9, \beta = 20$ とした。

この cepstrum lifter を用いて同一の資料を処理したスペクトルの例を図 9に、スペクトログラムの例を図 10に示す。図 3で顕著であった 2 次構造が取り除かれていることが分かる。

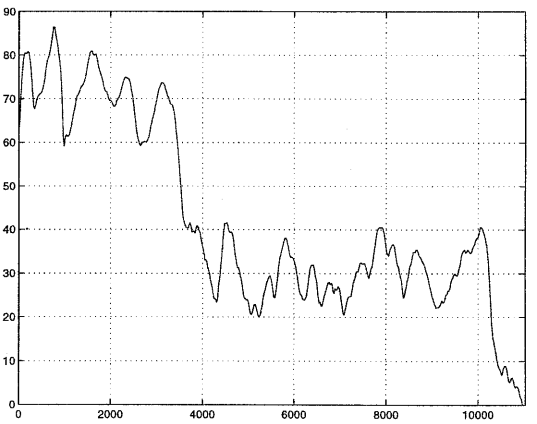

図 9: 男性の発声した "right" の STRAIGHT-core による処理結果の 330ms の時点における平滑化パワースペクトルから 2 次構造抑圧用の cepstrum lifter を用いて処理したパワースペクトル。

4 音源情報の精密化

STRAIGHT においては、非周期成分の時間分解能を広帯域信号の時間変化の知覚における分解能と同程度（約 3ms）に設定している。比較的長い時間窓により求められる周期成分と、このように短い時間窓により求められる非周期成分の配分を、主として基本周波数成分の周期性の強さのみに依存して切り替えていたことが前述の品質劣化につながったと考えられる。音声信号の周期性は、周波数帯域毎に異なっているため、この問題を本質的に解消するためには、multiband excitation Vocoder[3] のように、

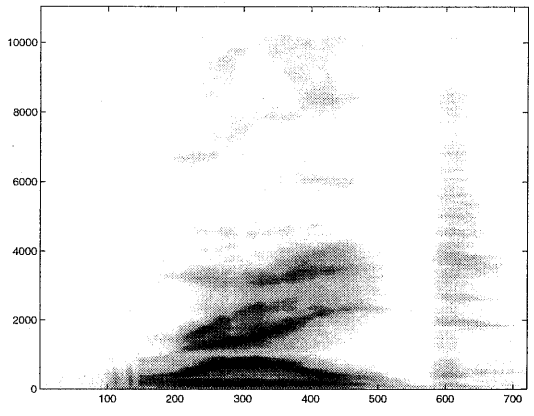

図 10: 男性の発声した “right” の STRAIGHT-core による処理結果から、提案した方法を用いて 2 次的な構造を取り除いた結果。

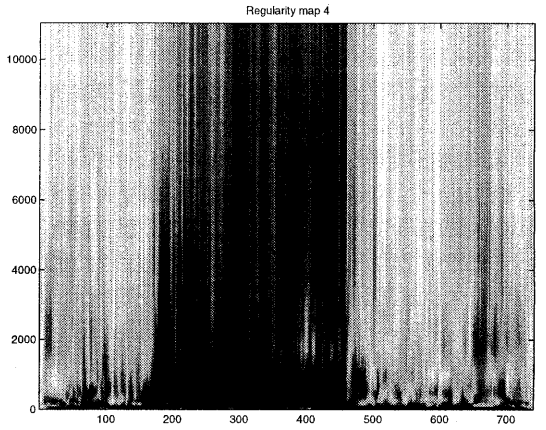

図 11: 補間された相関 (男声による wright の発音)。

周波数帯域毎の周期性に基づいた処理を導入することが必要である。

STRAIGHT のように基本周期に比較して短かな時間窓を用いた分析において局所的なスペクトルのみから周期性を評価することは困難である。また、声帯振動が厳密には周期的でないこと、声道形状の時間変化に伴い、インパルス応答が時々刻々変化して行くこと、声門の開閉の周辺で、気息性の雑音が生ずるために駆動波形がランダムに変動すること等により、基本周波数の情報を利用した comb filter を用いても、基本周期に同期して変動する成分を完全に除去することはできない。

ここでは、互いに重複する帯域幅 1/2 オクターブの Gabor 関数の時間差分をフィルタのインパルス応答として用い、帯域毎に基本周期だけ離れた位置の波形の複素相関の絶対値および同様な波形の包絡の相関に基づいて周期性を評価することとした。複素相関の絶対値を用いることにより、基本周波数の誤差および波形振幅の変動に耐性のある周期性の評価が可能となる。

音声の再合成には、周期成分と非周期成分それぞれに対応する 2 種類のスペクトログラムを用いる。今回の実験では、非周期成分に対応するスペクトログラムを 3ms の時間窓による短時間 FFT の結果により代用している。これらのスペクトログラムに基づいて、音声の周期成分と非周期成分とを別々に生成し、それらを、前述の帯域毎の複素相関の絶対値に基づいて混合することにより、急峻な閾値処理を回避することとした。それぞれの帯域の周期性は互いに重複するフィルタで計算されているため、合成に際しては、本来はそれらの重複を逆特性で補正した補間関数を用いる必要がある。しかし、予備実験の結果、逆に品質の劣化が生じたため、ここでは、逆特性を利用せず、付録に示す関数により補間した加重を用いることとした。図に同じ例について、補間された相関を示す。暗い部分が相関の高い部分を示している。この補間された相関ρから波形の混合比 $B(\rho)$ の導出には、次式で表わされる滑らかないき値関数を用いた。

$$B(\rho) = \frac{1}{1 + e^{-\alpha(\rho - \theta)}} \tag{2}$$

なお、全体の処理の流れと用いた加重を付録に示す。

ここでは、表示の便宜ために図 10と非周期成分に対応するスペクトログラムとを最小分散基準により合成して図 12に示す。

5 今後の課題

今回の報告では、少数の事例の詳しい解析に基づいて品質劣化に結び付く問題点を抽出し、対策を提案して来た。これらの対策が一般的に適用可能であるか、適用が常に品質の向上に結びつくのかは、多様な音声を用いたパラメタの最適化と評価により検証して行かなければならない。

ところで、本報告で紹介したスペクトルの 2 次構造は、声帯の振動と声道の相互作用についての興味深い問題を提起しているのか m 知れない。誉田らに

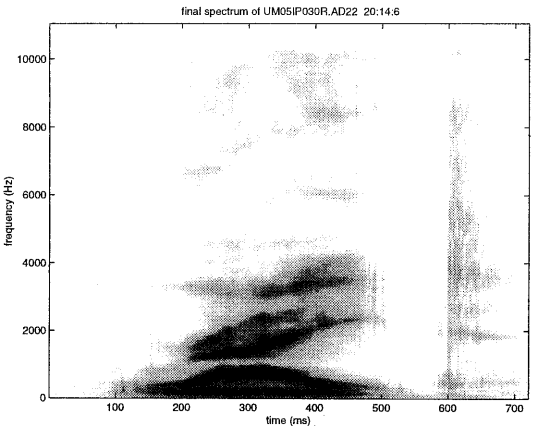

図 12: 男性の発声した "right" の STRAIGHT-core の結果に全ての対策を施して計算したスペクトログラム。

よる位相等化音源に関連して、「主要なピッチパルスの中間に小さなパルスを加えた方が音質が良くなる」現象が紹介されたことがある。今回紹介した資料以外でも STRAIGHT-core の処理で同様な 2 次構造が見える音声の例は多い。しかし、音源と声道の線形の相互作用のみでは、このような 2 次構造が比較的安定して認められることを説明することは困難である。紹介した例に見るような 20 次以上の調波成分にまで 2 次構造が観察されるためには、励振の位置が主要な励振の位置の間の中央から 5%以内に無ければならず、何らかの拘束が働いていることを想像させる。

6 まとめ

STRAIGHT における品質劣化要因を事例に基づいて紹介し、それらの問題点を解消することを目的として導入した、周期／非周期成分の処理の分離と加重合成、およびスペクトルの 2 次的構造の除去について説明した。これらの改良は、スペクトルの大局的な性質と時間的な性質とを、より操作し易い形に分離するため、音声知覚研究や音声の可視化ツールとしての STRAIGHT の有用性を高めるものでもある。

予備的な聴取は、本報告で提案した処理による品質改善効果を示唆しているが、これまでに提案してきた処理と競合する部分もあるため、本格的な聴取実験によるパラメタの最適化が必要である。また、STRAIGHT を様々な問題に応用する上で、処理量と記憶容量がボトルネックとなっているため、効率的なアルゴリズムの研究を並行して進める必要がある。

謝辞

STRAIGHT の品質に関しての指摘と試料の提供を頂いた ATR 人間情報通信研究所 山田玲子 博士に感謝します。本研究は、科学技術振興事業団による戦略的基礎研究推進事業 CREST による援助を受けて行われた。

参考文献

[1] 丁文, 藤澤譲, Nick Campbell, 樋口宜男, 河原英紀. STRAIGHT を用いた CHATR の韻律制御. 日本音響学会講演論文集, No. 1-2-8, pp. 211–212, October 1997.

[2] H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, Vol. 11, No. 2, pp. 169–177, 1939.

[3] Daniel W. Griffin and Jae S. Lim. Mutiband excitation vocoder. *IEEE Trans. ASSP*, Vol. 36, pp. 1223–1235, 1988.

[4] 河原英紀, 増田郁代. 時間周波数領域での補間を用いた音声の変換について. 信学技報, Vol. EA96-28, , August 1996.8.

[5] 河原英紀. 聴覚の情景分析と高品質音声分析変換合成法 STRAIGHT. 日本音響学会講演論文集, No. 1-2-1, pp. 189–192, October 1997.

[6] 河原英紀. 音声中のバースト雑音の検出能力について. 聴覚研究会資料, No. EA86-39, October 1986.

[7] Hideki Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proceedings of IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 2, pp. 1303–1306, Muenich, 1997.

[8] 東山恵祐, 陸金林, 中村哲, 鹿野清宏, 河原英紀. 4khz 帯域の STRAIGHT の品質評価と情報圧縮について. 日本音響学会講演論文集, No. 1-2-6, pp. 207–208, October 1997.

[9] 河原英紀, Alain de Cheveigné. 原理的に抽出誤りの存在しないピッチ抽出方法とその評価について. 信学技報, Vol. SP96-96, pp. 9–18, 1997.

[10] 河原英紀, 増田郁代. 音声分析・変換・合成法 STRAIGHT のスペクトル近似特性の評価と改良について. 信学技報, Vol. SP96-97, pp. 19–24, 1997.

[11] 河原英紀, 増田郁代, 東山恵祐. 音声分析・変換・合成方法 STRAIGHT-TEMPO における相補的な時間窓の利用について. 信学技報, Vol. SP97-32, pp. 21–28, 1997.

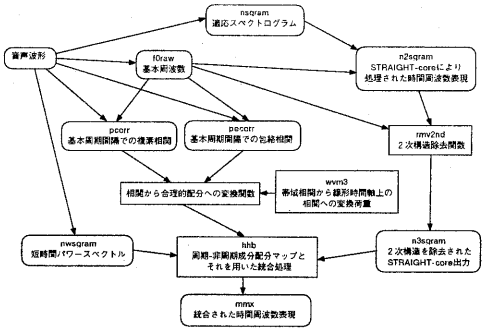

図 13: 精密化した分析処理の流れ。ここでは、2 次構造除去から直接表示のためのスペクトログラムを構成している。合成の場合には、周期信号成分については、9 月の研究会で提案した時間領域の処理をもちいた過剰平滑化の補償を行ったものを用いなければならない。この場合でも、補償の程度を理論的に求められたものの 2 倍にした方が音が良くなるように感じられる。

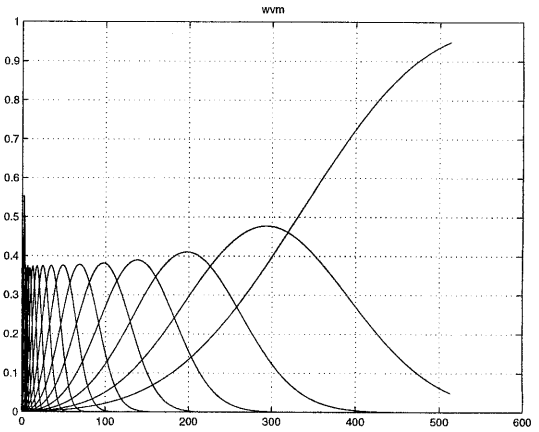

図 14: Gabor 関数を直接用いた場合の補間関数。

[12] 河原英紀, 津崎実, Roy D. Patterson. 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. 音響学会聴覚研究会, Vol. H-96-79, pp. 1–8, 1996.11.

[13] 河原英紀, 山田玲子, 久保理恵子. STRAIGHT を用いた音声パラメタの操作による印象の変化について. 聴覚研究会資料, Vol. H-97-63, , 1997.9.

[14] 久保理恵子, 山田玲子, 河原英紀. STRAIGHT による分析合成音声を用いた外国語音声知覚訓練. 聴覚研究会資料, Vol. H-97-64, , 1997.9.

A 処理の流れ

処理の流れを整理して図示する。

B 周期成分と非周期成分の合成に用いた加重

対数周波数軸上で連続的で滑らかな cos を 1/2 あるいは、2/3 づつ重複するように配置して周波数方向での平滑化の効果がでるように設計した。具体的にな補間関数の形を以下の図に示す。差分信号に対して Gabor 関数を用いてフィルタリングしたため、Gabor 関数の中心周波数と実効的なフィルタの中心周波数は一致していない。 補間関数は、全てを加算したときに各帯域の荷重が 1 となるように正規化した。

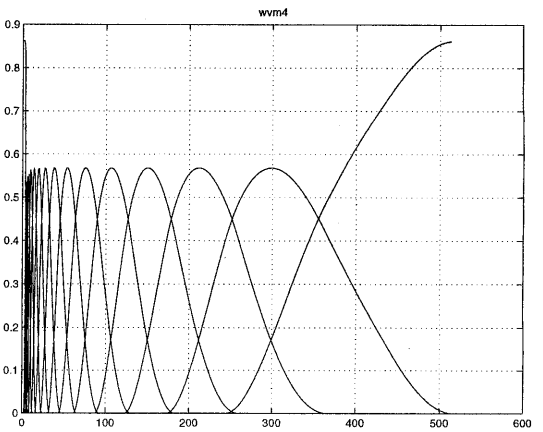

図 15: 2/3 重複の cos（位置補正有り）を用いた補間関数。