非公開

TR-H-206

0024

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

河原　英紀，Alain de Cheveigné

1996.11.26

ATR人間情報通信研究所

〒619-02　京都府相楽群精華町光台2-2　TEL: 0774-95-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Telephone: +81-774-95-1011

Fax : +81-774-95-1008

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

Error Free F0 Extraction Method and Its Evaluation

河原 英紀、 Alain de Cheveigné
第一研究室 *kawahara@hip.atr.co.jp*

平成8年 11月 25日

Abstract

あらまし　基本波成分の瞬時周波数の計算に基づく音声の基本周波数の新しい抽出方法を提案し、発声と同時に記録したEGG信号との相互比較と人工的信号を用いたシミュレーションにより、性能の評価を行なった。基本周波数の探索範囲を40Hz～800Hzとした場合、後処理無しの状態で既に本方法の性能は、従来の方法の性能を凌駕している。因に、女性の発声した100文章音声の分析結果は、全分析結果の50%以上が、EGGの分析結果の±0.3%以内に入っていることを示した。基本周波数および基本周期に対して等方的なGabor関数を用いたwavelet分析に基づいて新たに『基本波らしさ』の指標を定義することにより、基本周波数を抽出せずに基本波成分を選択できるようにしたことが本方法の鍵となっている。(平成8年 11月 25日 5:46 P.M. 版、音声研究会(1997.1.17)用は、本資料から抜粋)

Abstract　A new F0 (fundamental frequency) extraction algorithm is proposed, which does not introduce extraction error in principle. The key is to define F0 as the instantaneous frequency of the fundamental component of the signal. This seemingly contradictory definition is made practical by introducing a new 'fundamentalness' measure. The measure is defined based on wavelet analysis using an iso-metric Gabor function. A series of evaluation using a database of simultaneously recorded sentence materials with EGG (Electro Glott-Graph) signals was conducted. The performance of the proposed method outperformed conventional methods without any post-processing for a 40Hz to 800Hz F0 search range. For example, over 50% of analysis freme of a 100 sentence female speech databese showed that the agreement is within ±0.3%.

1 はじめに

筆者らは、これまで、人間の音声コミュニケーションにおける音声知覚と生成の相互作用の研究を行なってきた[12]。その結果、音声の基本周波数の制御における、生成と知覚との間の相互作用の定量的測定に初めて成功し、興味深い二重構造の制御機構を見い出した。しかし、それらの研究においては、既存の基本周波数抽出プログラムを使用していたため、8Hz 以上の周波数領域での現象が本来の発声・知覚機構によるものなのか抽出方法の副作用によるものであるかを明らかにすることができなかった。

一方、知覚と生成の相互作用の研究をピッチ知覚以外の領域に拡張することを目的とした研究の必要に迫られて、VOCODER の原理そのものでありながら品質が非常に良い分析・変換・再合成法 (STRAIGHT: Speech Transformation and Representation based on Adaptive Interpolation of Gaussian weigHTed spectrogram)[1]を発明した[10]。この方法は基本周波数の情報を積極的に用いているため、基本周波数抽出法の良否が直接変換音声の品質を左右する。従来の基本周波数の定義に基づく様々な抽出方法を試用しても[7, 6, 4]、wavelet の位相に注目した時間分解能の高い方法を用いて更に視察による修正を丁寧に加えても[9]、細部のジッタや有声／無声の判定の影響か、元の音声と比較するとわずかな濁りが感じられた。

そこで、これらの問題点を解決することを狙い、改めて音声変換のための基本周波数の抽出とは何であるかという問題設定自体を根本的に見直すこととした。その結果、以下に報告するように、比較的簡単な処理で基本波成分の瞬時周波数を求めて基本周波数とすることで、従来の方法を凌駕する性能を有するアルゴリズムを作成することができた。

以下、順を追って説明する。まず、最初にこれまでの方法で用いられてきた周期信号の定義と、本方法で用いる信号の定義を説明する。本方法の定義は、音声等の時間的に変化する信号に適した表現として調波的構造を有しかつそれぞれが AM-FM 変調されている信号として定義している。この信号に基づいて、基本周波数の抽出が最低次の成分の瞬時周波数の抽出の問題であることを説明し、本方法の中心的な概念である『基本波らしさ』を提案し、具体的な定義を導く。次に、実装の節において、『基本波らしさ』を用いて具体的な計算方法を明らかにする。これらの準備の下で、まず周期的なパルス列を用いたシミュレーションにより、本方法の精度と耐雑音性を調べる。次に、実際の音声の分析例を示し、方法の特徴について説明する。その後、ATR 音声翻訳通信研究所の Campbell らによって整備された多量の音声データと同時記録された EGG からなるデータベースのサブセットを用いた実験による従来の方法との比較結果を示し、包絡情報を用いる方法についての予備試験の結果を示す。今後の課題に関する節では、本方法が聴覚におけるピッチ知覚の計算論レベルでの議論を与える可能性について議論し、最後に一般の生体信号の解析への応用として脈波データの解析結果を紹介する。

なお、本方法では、『基本波らしさ』を「AM 変動や FM 変動が最も少ないこと」として定義していることが本質的な役割を果たしている。本方法は、この手がかりに基づいて wavelet 変換の位相の時間微分として定義される瞬時周波数を求めており、基本的には時間領域の方法である。そこで、本方法の略称として TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator) を用いることを提案する。

[1]GHT に対応する英語が最初の提案とは異なっている。これは、本報告および同じ研究会で発表されるもう一つの報告で Gaussian の重要性が明らかになったからである。あるいは、S も Spline based とした方が良いかも知れない。

2 F0、ピッチ、瞬時周波数

従来の方法では、周期信号の定義に基づいて、以下で定義される周期Tを求め、その逆数を基本周波数としていた。分析の対象とする周期信号を$p(t)$とし、$n \in N$を任意の整数とする。

$$p(t) = p(t + nT) \tag{1}$$

あるいは、このような信号の周波数領域での表現を用いて、次に示すようなFourier変換により最も良く近似できるようなパラメタの値f_0として基本周波数を求めることも行なわれていた。

$$p(t) = \sum_{k \in N} \alpha_k \sin(2\pi k f_0 t + \phi_k) \tag{2}$$

一方では、人間のピッチ[2]感覚についての実験心理学的知見に基づいて、聴覚モデルの出力から、いかにそれらの実験結果と良く整合する値を計算するかを追及する研究も行われていた[6]。

STRAIGHTのような高品質の分析・変換・合成系の音源パラメタとしての基本周波数を求めようとする場合、抽出方法の理論的背景の中に、中途半端な近似や過剰な時間平滑化処理が含まれていたり、実装にアドホックな後処理等が含まれていると、それらは明らかな品質の劣化として耳につく。そのような意味で、音声のように常にスペクトルや駆動音源の周期が変化するような信号では本来成立していない数学的な周期性の定義に基づく方法は不適切である。本報告では、時間とともに振幅や速度が変化する声帯の振動により主に駆動される、時間とともに変化する伝達特性の影響を受けた信号として音声を捉える。この性質をできるだけ忠実に表現するように定義した以下のようなAM-FM信号を分析対象として議論を展開する。

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin\left(\int_{t_0}^{t} k(\omega(\tau) + \omega_k(\tau))d\tau + \phi_k\right) \tag{3}$$

$\alpha_k(t)$は、振幅変動(AM: Ampiltude Modulation)項を表わし、$\omega(\tau)+\omega_k(\tau)$は、周波数変動(FM: Frequency Modulation)項を表わす。ここで、$\omega(\tau) \gg \omega_k(\tau)$を要請する。すなわち、音声信号は、各周波数帯域毎に少しづつ異なった基本周波数を有すると考えるのである。この式は、McAulayらによるSinusoidal Representation[15]と類似しているが、ほぼ周期的な信号を表現する構造を持ち込んでいることが大きく異なっている。このような構造は、我々の発明したSTRAIGHTという音声分析・変換・合成の枠組みと組み合わせることにより、初めて自然な音声の合成に結び付けることができる。正弦波を直接正確に求めようとすることは、逆に人間の聴覚系の情景分析能力に対する大きな妨害を生みだし、品質の劣化に結び付く。

本方法の定義で用いたほぼ調波構造を保ちながらもそれぞれの周波数帯域において少しづつ異なった基本周波数を有するというモデルは、声門の閉止が声帯の振動によって基本的には規定されながらも、声帯の表面を伝播する微小な波動により規則的なタイミングとは微小に前後して生じる状況をモデル化していると見ることもできる。実際、後で見るように本

[2]音声信号処理の分野での従来の用語との整合性を保つために、本報告の題名は「ピッチ抽出」という言葉を基本周波数の抽出（より正確には基本波の瞬時周波数の抽出）の意味で用いた。このような用法は、心理量であるピッチをあたかも物理量であるかのように見なすという混乱を招くため望ましくない。以下では、「ピッチ」は心理量に言及する場合にのみ用いることとする。

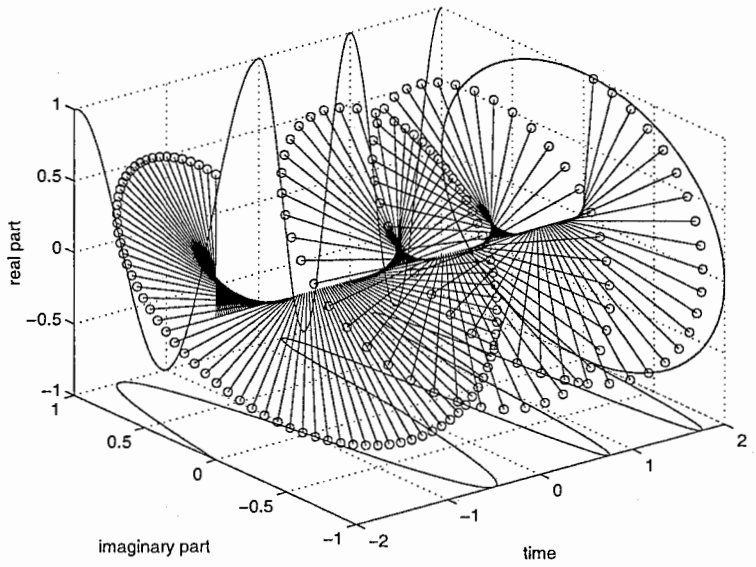

図 1: 瞬時周波数の説明図。

方法への入力を帯域フィルタ出力の包絡信号とすれば、それらの周波数帯域毎に異なった基本周波数を求めることができる。

STRAIGHT のオールパスフィルタによる音源の実装方法は、この分析方法で得られるような周波数帯域毎にわずかに異なった基本周波数を有する駆動信号を実現することができる。ただし、STRAIGHT のオールパスフィルタによる実装の場合、長期的な平均値は全ての周波数帯域で同一になるという拘束条件が課される。実は、この条件は実際の声帯音源にも存在する拘束であるため、実質的には制限事項とはならない。

この定義に基づけば、$k = 1$ に対応する成分の瞬時周波数を以下のようにして計算することにより、基本周波数$\omega(t)$ が求められる。

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{4}$$

$$\phi(t) = \arctan\left[-\frac{H[s(t)]}{s(t)}\right] \tag{5}$$

ここで $H[\,]$ は信号の Hilbert 変換を表わす。また、位相$\phi(t)$ は、$\pm 2\pi$のジャンプが取り除かれた、連続的な信号として表わされているものとする。

直観的ではあるが、上記の式の意味を説明する模式図を図 1に示す。この図では、信号 $s(t) = \cos(2\pi\omega\eta t)$（ただし、$\eta = 1\ (t \leq 0), \eta = 2\ (t > 0)$）を実部に、その Hilbert 変換 $H[s(t)]$ を虚部として、一定時間毎の値を小円とスポークによって表わしている。瞬時周波数は、スポークの回転速度に相当する。図では、途中で瞬時周波数が二倍になっている様子が、スポークの密度の変化として読み取られる。

瞬時周波数を基本周波数の抽出に用いようとする試みには幾つかの先行研究があり、優れた性能も報告されている [1, 2]。しかし、それらの報告では、従来の基本周波数の定義を問い直すまでには踏み込んでいなかった。本報告では、従来のように周期信号の定義を音声のような本来時間的に変化し続ける信号の基本周波数の定義として用いるのは、誤りであると主張する。その代わりに、時間的に変化する調波類似の構造を持つ複合 AM-FM 信号と

して音声を表わし、後で定義する『基本波らしさ』が最大の部分のチャンネルの出力の瞬時周波数を以て基本周波数とすべきであると提案する。なお、瞬時周波数は信号が周期的な場合には、従来の定義から計算される基本周波数と一致する。

本方法の拡張として、入力に信号そのものを用いるのではなく、信号の包絡あるいはフィルタ処理された信号の包絡を用いることができる。このような拡張により、本方法は、missing fundamental として知られる状況においても基本周波数を抽出することができるようになる。実際の音声分析の部分で具体例を紹介する。

2.1 『基本波らしさ』の備えるべき性質

前節の議論により、基本波の瞬時周波数を求めれば基本周波数が求められることが分かった。しかし、このままでは、基本波を求めるためには基本周波数の情報が必要となり循環論法に陥ってしまう。そこで、以下のようにして『基本波らしさ』を周波数の情報を直接用いずに表わす指標を提案することにより、この循環を断ち切ることとする。

基本波は、音声のように繰り返し生ずるイベントで駆動される信号の場合、イベントの発生速度に対応する周波数の波を表わす。音声の場合は、声帯の振動周波数の波が基本波となる。音声では、この声帯の振動が声門の開閉を通じて呼気流を変調する。低い周波数領域では、声門の面積そのものが、高い周波数領域では、面積変化波形上の特異点の性質が、声道に対する励振を規定する。しかし、これらは共通の声帯の振動という現象の現われであり、Bregman[3] の言うように、それぞれの調波成分間の時間的変動は類似することとなる。

したがって、このような共通の変動を取り除いた場合に、基本波においてだけ変動が少なくなり、その他の調波では変動が増加するような正規化された量を見い出し、その逆数あるいは符号を反転した量を『基本波らしさ』であると定義すれば良いことが分かる。上の議論から、もし基本波に AM や FM がかかっていたとしても、全ての調波成分がそれを共有するので、調波間での『基本波らしさ』の順序は変わらない。言葉を変えると、基本波以外の調波において、それらの基本波と共通する AM や FM 成分に加えて、基本波でないことによる調波の相互干渉で AM や FM が生ずるようにすれば、常に基本波において『基本波らしさ』が最大となる。STRAIGHT で用いていた特別な Gauss 窓の性質を利用すれば、この要請を容易に満たすことができる。

2.2 Gabor 関数を用いた『基本波らしさ』の定義

次のように定義される Gabor 関数から求められる関数 $g_{AG\tau_0}(t)$ を分析 wavelet とした wavelet 分析 [5] を行なう。この関数を用いれば、基本波に相当する成分においては、それ以上の成分からの干渉を受けず、基本波以外の調波成分においては、調波相互が干渉するという条件が実現される。

$$g_{AG\tau_0}(t) = g_{\tau_0}(t-\tau_0/4) - g_{\tau_0}(t+\tau_0/4) \quad (6)$$

$$g_{\tau_0}(t) = e^{-\pi(t/\tau_0)^2} e^{-j\frac{2\pi t}{\tau_0}} \quad (7)$$

$g_{\tau_0}(t)$ は wavelet 変換の許容条件を満たさないが、それらを $\tau_0/2$ だけ離して差を取った $g_{AG\tau_0}(t)$ は、許容条件を満たす[3]。これは、第二調波からの干渉だけを除去することを意味する。な

[3]実用上は、等分解能の Gabor 関数を出発点とするよりも、時間方向に少し (1.3 倍程度) 広げた関数を用

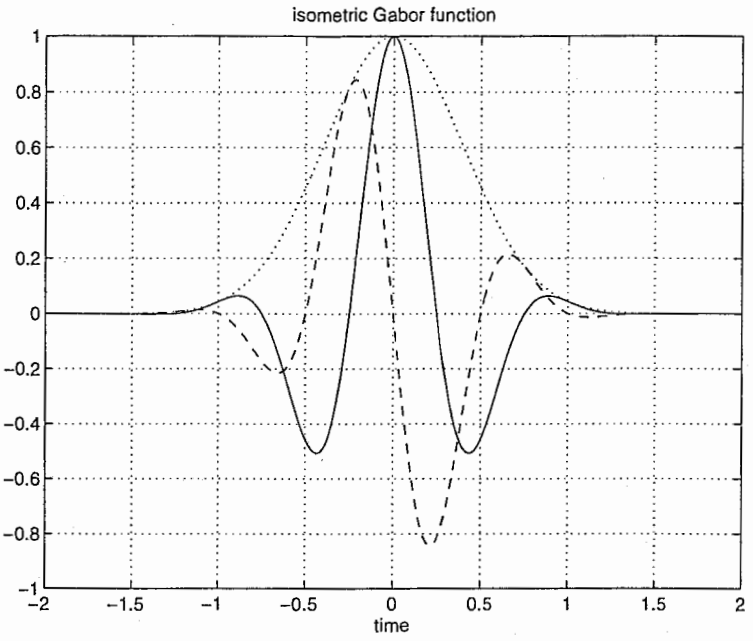

図 2: 時間分解能と周波数分解能がバランスした Gabor 関数の時間波形。

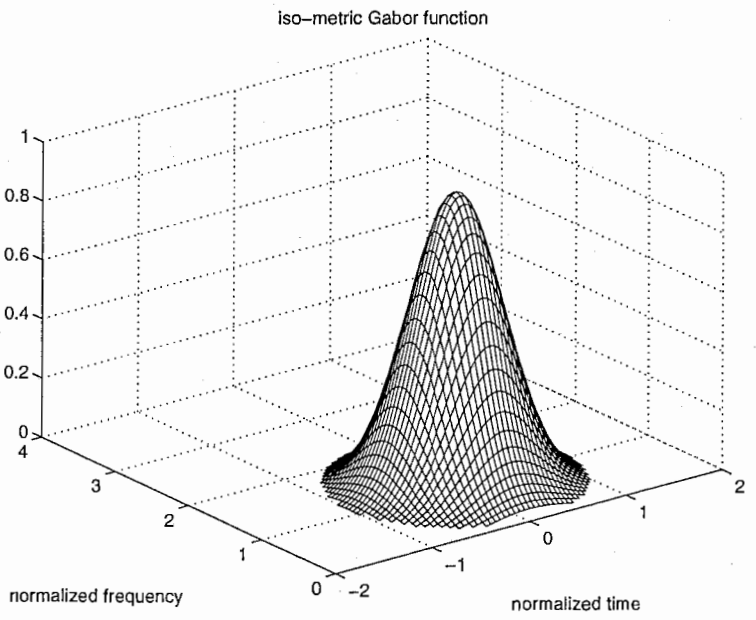

図 3: 時間分解能と周波数分解能がバランスした Gabor 関数の時間周波数表現。

ぜなら第三調波からの干渉は−100dB 以下の大きさになるので、実用上無視できるからである。

いた方が良い。厳密な議論は別に行なう必要があるが、周波数の移動が大きい場合、オクターブに置くチャンネルの数が十分に確保できない場合に有効である。

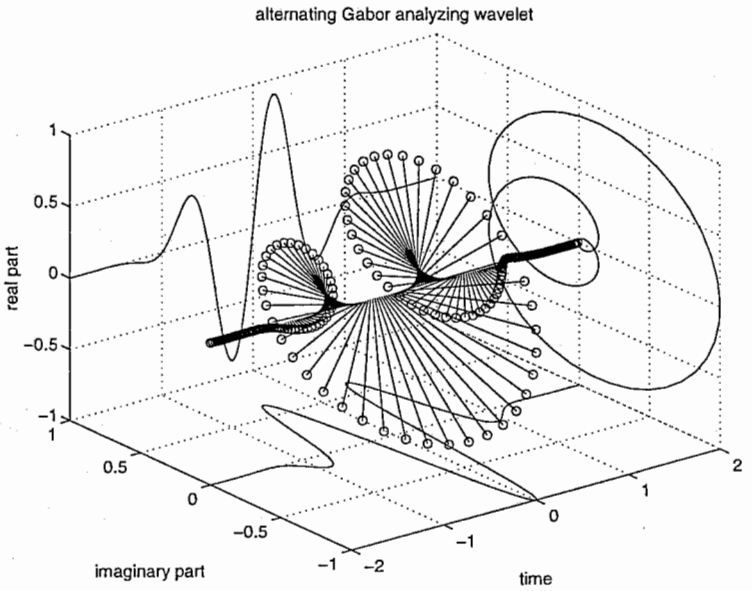

図 4: 時間分解能と周波数分解能がバランスした Gabor 関数から作成された wavelet の時間波形。

この $g_{AG\tau_0}(t)$ を作成するために用いている Gabor 関数 $g_{\tau_0}(t)$ は、

$$g_\alpha(t) = \frac{1}{2\sqrt{\pi\alpha}} e^{\frac{t^2}{4\alpha}} \tag{8}$$

として定義される Gabor 関数で、$\alpha = \tau_0^2/4\pi$と置いた場合に相当する。図 2に$\tau_0 = 1$ とした時の時間波形を、図 3に時間周波数表現の絶対値を示す。図中では基本周波数ならびに基本周期の間隔は 1 となるように正規化されている。この尺度において、$g_{\tau_0}(t)$ は等方的であることが分かる。

この Gabor 関数から作成された $g_{AG\tau_0}(t)$ の図 1 と同様の表示を図 4に示す。Gabor 関数の実部と虚部が入れ替わり、ややねじれが強くなった形をしている。

この $g_{AG_1}(t)$ を用いた wavelet 変換 $D(t,\tau_0)$ を以下のように表わすことにする。

$$D(t,\tau_0) = |\tau_0|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t)\overline{g_{AG_1}\left(\frac{t-u}{\tau_0}\right)} du \tag{9}$$

実際には $g_{AG\tau_0}(t)$ の振幅は指数関数のオーダーで急速に減少する。したがって積分の範囲からは $g_{AG\tau_0}(t)$ が実質的に零であるとみなすことのできる範囲を取り除くことができる。これは、実時間処理に都合の良い性質である。

この関数に基づいて、基本波らしさを表わす指標 $M(t,\tau_0)$ を以下のように定義する。

$$\begin{aligned} M &= -\log\left[\int_\Omega \left(\frac{d|D|}{du}\right)^2 du\right] + \log\left[\int_\Omega |D|^2 du\right] \\ &\quad -\log\left[\int_\Omega \left(\frac{d\arg(D)}{du}\right)^2 du\right] + 2\log\tau_0 + \log\Omega \end{aligned} \tag{10}$$

ここで、Ωは wavelet のサイズによって決まる積分の範囲を表わす。最初の二つの項は、相対的振幅変動速度（AM 速度）の自乗平均値の逆数の対数に相当する。第三の項は、瞬時周波数変動速度 (FM 速度) の自乗平均の逆数の対数に相当する。最後の項は、微分の値が対象とする信号の周波数によって変化することを正規化するための補正項である。

ここまでは時間分解能と周波数分解能がバランスした Gabor 関数から出発して議論を組み立てて来た。この関数をフィルタとして見ると、高い周波数側では急峻で、低い周波数側ではなだらかな聴覚フィルタに類似の形をしていることが分かる。このような形態のフィルタを用いて『基本波らしさ』を求めると、基本波以外の調波では、下側の調波からの干渉とフィルタの形状が調波間隔と比較して相対的に広がることによる調波間の干渉により M の値は低下する。その結果、このような性質を持つ広いクラスのフィルタを用いた場合においても、基本波において、M は、最大となる。すなわち、組織的に配置された τ_0 の様々な値に対応するフィルタ出力についての M を計算し、それらの中で最大の M を与える τ_0 を選ぶと、それが基本波成分の周波数に対応することになる。なお、この指標は無次元化されている。したがって、M の値そのものは、一般的な『基本波らしさ』の程度を表わすものと見なすことができる。

こうして求められる τ_0 は、比較的なだらかな関数の頂点付近の位置を示す指標なので、雑音などの存在に敏感となり、そのまま基本周波数の計算に用いることは適切ではない。実際には、この τ_0 によってフィルタを選択し、そのフィルタ出力 $D(t,\tau_0)$ から瞬時周波数 $f(t)$ を

$$f(t) = \frac{1}{2\pi}\frac{d\arg\left(D(t,\tau_0)\right)}{dt} \tag{11}$$

によって求めて、基本周波数の値とする。

以上の議論と同型の議論は、Gabor 関数以外を出発点の関数として用いても展開することができる。そのようなものの中で興味深いのは、聴覚の計算理論として入野 [8] が提案している Gammachirp を出発点とする議論である。

3 実装

実際の処理のデータの流れを図 5に示す。本方法は、基本的に後処理や探索を含まないので、演算速度さえ十分にあれば、簡単にハード化でき実時間処理ができる。wavelet 変換は、図 5では 2 重に書いてあるが、これは同一の分析を再利用することができる。以下、簡単に処理の流れを説明する。本方法は、どのような周波数領域でも扱うことができるため、生体信号等の一般の信号処理に用いることもできる。しかし、ここでは音声の基本周波数の抽出を例として、具体的なパラメタの設定も含めて説明する。

[Step:1] 図 5において、入力された音声波形は、Gabor 関数から求められた一組のフィルタ群（具体的には、40Hz から 800Hz までの 1/12 オクターブ毎の 52 個のフィルタ）により分析され、52 個のチャンネルによる wavelet 変換の値が並列に求められる。

[Step:2] 各チャンネルの wavelet 変換を用いて、チャンネル毎に『基本波らしさ』が求められる。これらの値は、標本化周波数のレートで求めることができる。実際には、適当なダウンサンプリングと組み合わせて、例えば STRAIGHT で用いる場合には、1ms 毎の値を求める。こうして求めた『基本波らしさ』が最大となるフィルタ番号を選択する。

[Step:3] 選択されたフィルタに対応する wavelet 変換を用いて、瞬時周波数を求め、これを基本周波数の第一近似値とする。

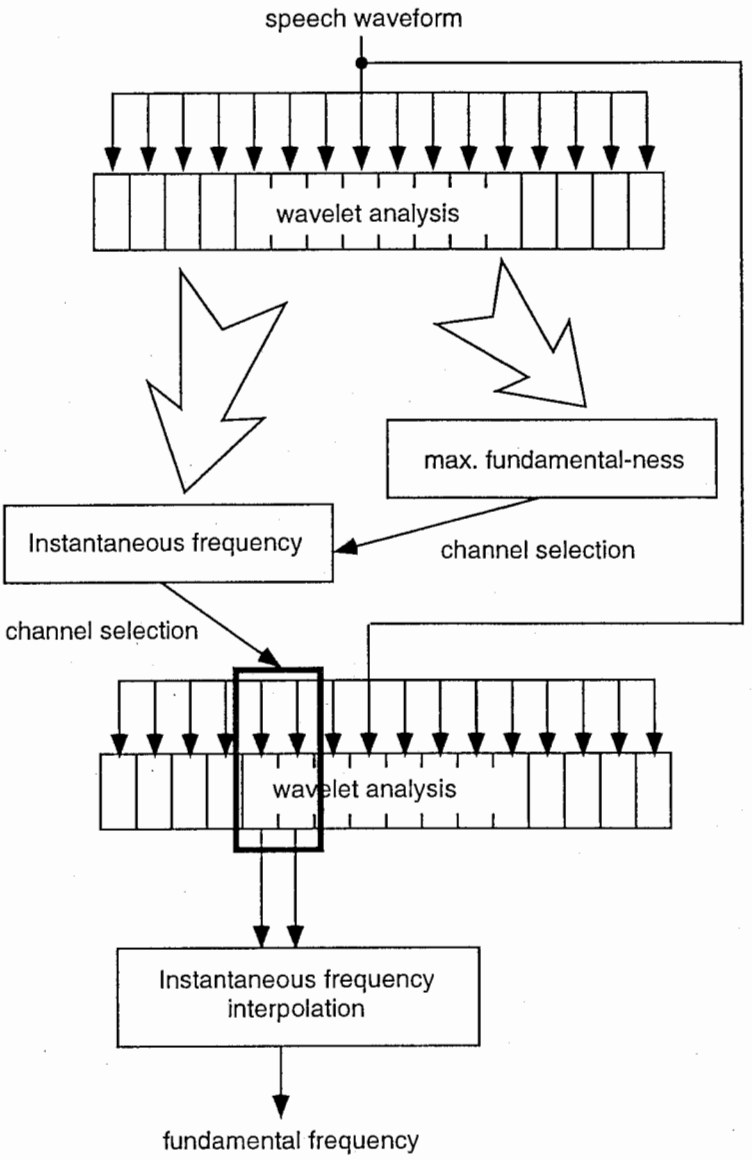

図 5: 基本周波数抽出処理のデータの流れ。

[Step:4] 求められた基本周波数の第一近似値からその基本周波数に最も近い中心周波数を有するフィルタを二つ選択する。基本周波数の第一近似値は、二つのフィルタの中心周波数の中間に位置する。

[Step:5] 基本周波数の第一近似値を挟む二つのフィルタのそれぞれの出力から瞬時周波数を求め、第一近似値の位置に基づいて、それらの値の加重平均として基本周波数の値を抽出する。具体的には、基本周波数の第一近似値を f_1、その上にあるフィルタから計算される瞬時周波数を f_u、その下にあるフィルタから計算される瞬時周波数を f_l、それぞれのフィルタの中心周波数を c_u, c_lとし、基本周波数 f_0を次式により計算する。

$$f_0 = f_l + (f_u - f_l)\left(\frac{f_1 - c_l}{c_u - c_l}\right) \tag{12}$$

以上説明してきた処理は、Matlab 上に実装されている。実装例のリストを付録として添

付する。例題として挙げた実装は、本方法の実現可能性を実証するためのものであり、各部分の処理パラメタを過剰品質気味に設定している。そのため、1秒の音声を処理して1ms毎に基本周波数の値を抽出するのに約200Mflopsを必要としている。しかしこの値は、実装とアルゴリズムの工夫で1/10程度にできるため、現在簡単に入手できるDSPを用いても実時間処理が可能であると見込まれる。

4 実験

提案した方法は、信号処理だけを用いており、推定に類する過程を含んでいない。この意味で、本方法は原理的に誤りを含まない基本周波数の抽出方法といえる。しかし、対象とする信号以外の妨害音（たとえば雑音）を含む場合、本方法は、雑音と目的音が合成された信号の中から最も基本波らしい部分を選び瞬時周波数を計算して基本周波数を抽出する。こうして抽出された基本周波数の値は、雑音の影響のため雑音の無い場合の値と異なる。

この雑音の影響は、式4の形を見ると、信号対雑音比に直接支配されることが分かる。また、『基本波らしさ』も信号対雑音比に同様に直接支配されることが分かる。従って、雑音の存在による基本周波数の抽出値のずれの範囲は、雑音が無い場合の基本周波数の値が分からなくとも、『基本波らしさ』の値から計算できることが分かる。これは、非常に有用な性質である。

以下の実験では、この性質を、シミュレーションならびに、EGGと同時記録した音声データベースにより確認する。

4.1 パルス列

ここでは、一定の周期(100Hz)を有するパルス列に60dBから0dBまでのS/Nとなるように系統的に雑音を加えて、『基本波らしさ』の指標と抽出された瞬時周波数と合成に用いた音源の周波数との食い違いを調べた。本方法は、標本化周波数ならびに基本波の周波数に依存しない汎用の方法であるので、他の基本周波数について調べる必要は無い。

まず、信号対雑音比と抽出成功率ならびに抽出された基本周波数の標準偏差との関係を表1に示す。抽出された基本周波数の測定点数は、各条件毎に800である。表中の抽出率は、合成に用いた値から10%以上離れた場合を除外した区間の数である。最後の(envelope)と記したものは、信号そのものではなく信号とそのHilbert変換をそれぞれ実部と虚部として作成した信号の絶対値（包絡信号）を信号波形とみなして本方法で分析した結果である。この手法を用いると、任意の周波数帯域に存在する信号の包絡を計算することによりベースバンドに変換し、包絡の変化の基本周波数を求めることができる。

図6に、『基本波らしさ』と抽出された基本周波数の偏差との散布図を示す。前述したように、『基本波らしさ』が偏差の存在範囲を明瞭に規定していることが分かる。『基本波らしさ』Mは、スケーリングしてdBを用いて表わしている。図6では、『基本波らしさ』が75以下の場合に、上の方に異常値が認められる。この部分では、パルス列ではなく雑音の部分の方がより『基本波らしい』性質を有していたことになる。

図7は、同じ結果から求めた『基本波らしさ』と標準偏差との関係である。この関係を利用すれば、抽出された基本周波数の標準偏差がある基準を満たす範囲だけを容易に選択することができる。例えば、『基本波らしさ』が80以上の部分だけを用いれば、基本周波数の

S/N ratio	% success	standard deviation
∞	100%	0.004 Hz
40 dB	100%	0.13 Hz
30 dB	100%	0.28 Hz
20 dB	100%	0.86 Hz
10 dB	95.7%	2.77 Hz
0 dB	43.0%	6.34 Hz
0 dB (envelope)	86.5%	5.22 Hz

表 1: パルス列の信号対雑音比と抽出された基本周波数の標準偏差

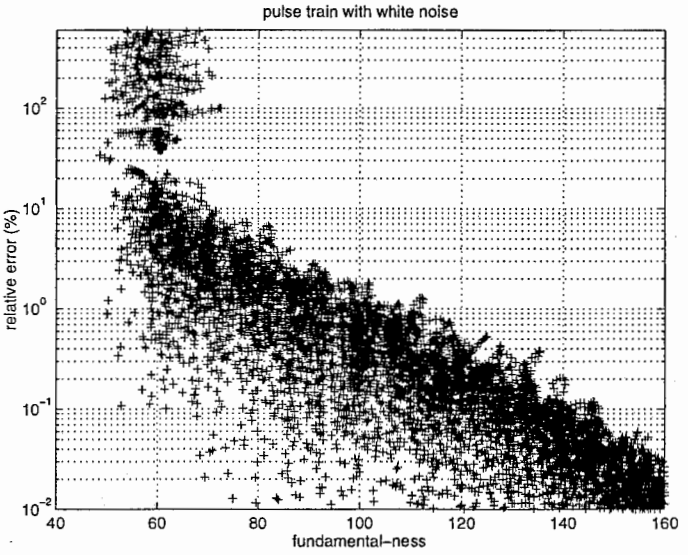

図 6: 求められた『基本波らしさ』の指標と誤差との散布図。

標準偏差が基本周波数の 2%程度となり、抽出結果は観測できない真の値の±6%の範囲内に 95%以上の確率で入っていることが期待できることとなる[4]。90dB が 1%の標準偏差に相当し、20 数値が増加すると標準偏差が 1/10 になると考えれば良い。

4.2 実音声の分析例

図 8に女性の発声した「交渉に追われています。」という音声を本方法により分析した結果を示す。図の最上段は、音声波形である。その次の段は、40Hz から 800Hz までの交番 Gabor フィルタ群で処理された信号のパワーの合計値である。その次の段は、抽出された基本周波数である。その下の段は、その基本周波数の成分の『基本波らしさ』を示す。その下の段は、抽出された基本波成分を含むチャンネル出力のパワーを示す。最後の段は、たて軸

[4]ここの議論は精密ではない。厳密には雑音の分布を仮定して議論する必要がある。このシミュレーションでは、無相関の正規雑音を用いた。

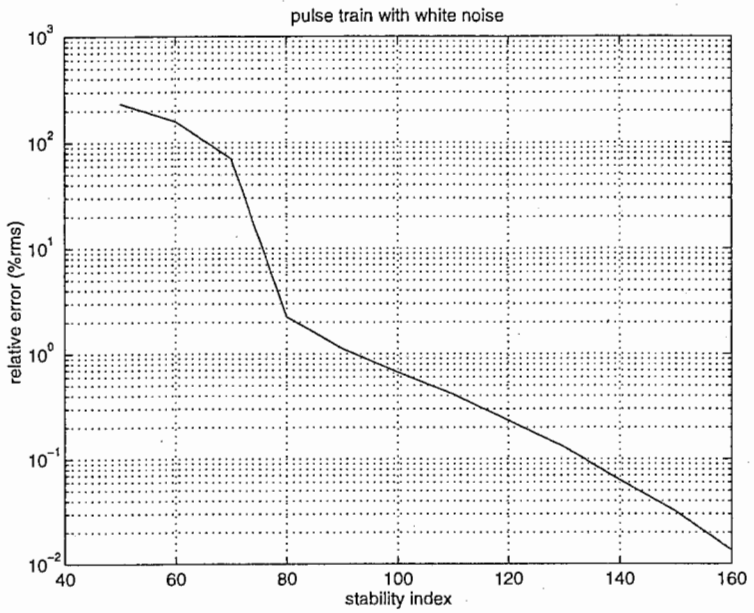

図 7: 求められた『基本波らしさ』の指標と対応する標準偏差。

をチャンネル番号、横軸を時間として『基本波らしさ』を濃度で表示したものである。『基本波らしさ』が強いほど濃い色で表示している。

図 9は、資料のオフセット印刷で濃度表示が見にくい場合のために、最初の 400ms の部分を 3 次元で表示したものである。『基本波らしさ』は、基本波が本当に存在する部分で際立って明瞭に大きな値を示している。この『基本波らしさ』自体、声帯の振動についての興味深い情報を豊かに与えてくれる。また、ここでは詳しく説明しないが、wavelet 変換の振幅成分表示も声帯振動や声門の開閉に伴う特異点の性質や配置についての情報を与えてくれる。

図 10に男性の発声した「爆音が銀世界の高原に広がる。」という音声を本方法により分析した結果を示す。/hirogaru/の部分では、子音の発声に伴なって基本周波数が一瞬低下する部分や、無声破裂子音の直後に高い基本周波数が急速に低下するマイクロプロソディーが明瞭に抽出されていることが分かる。

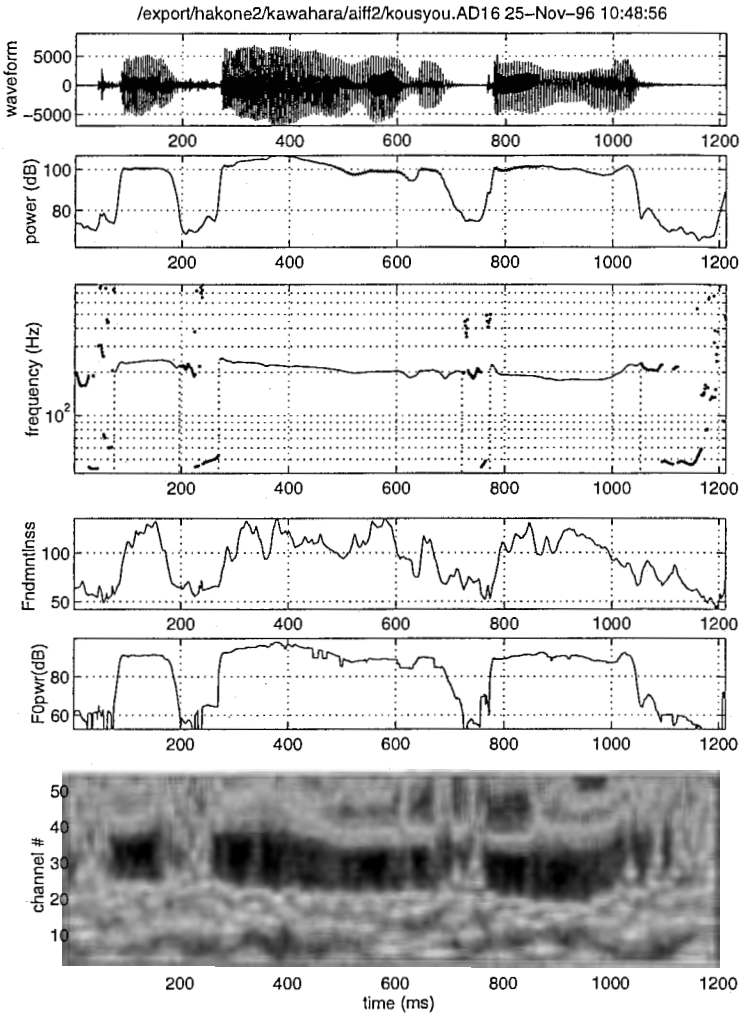

図 8: 女性の発声した音声「交渉に追われています。」の分析例。

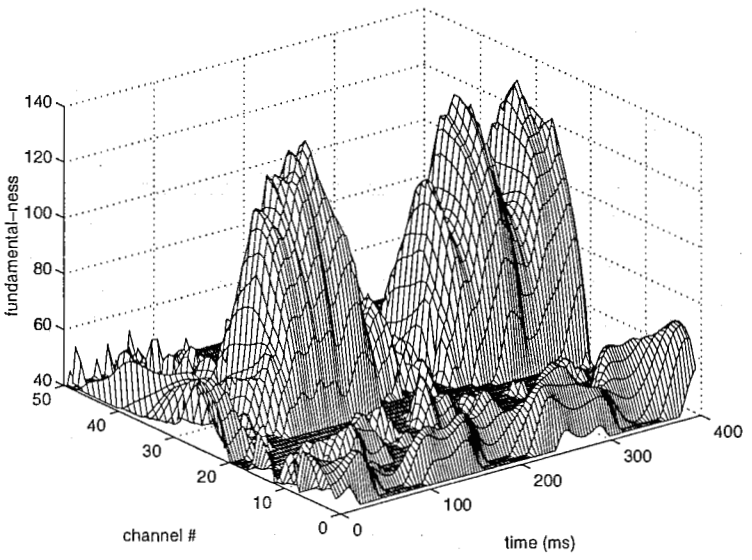

図 9: 女性の発声した音声「交渉に追われています。」の『基本波らしさ』の三次元表示。

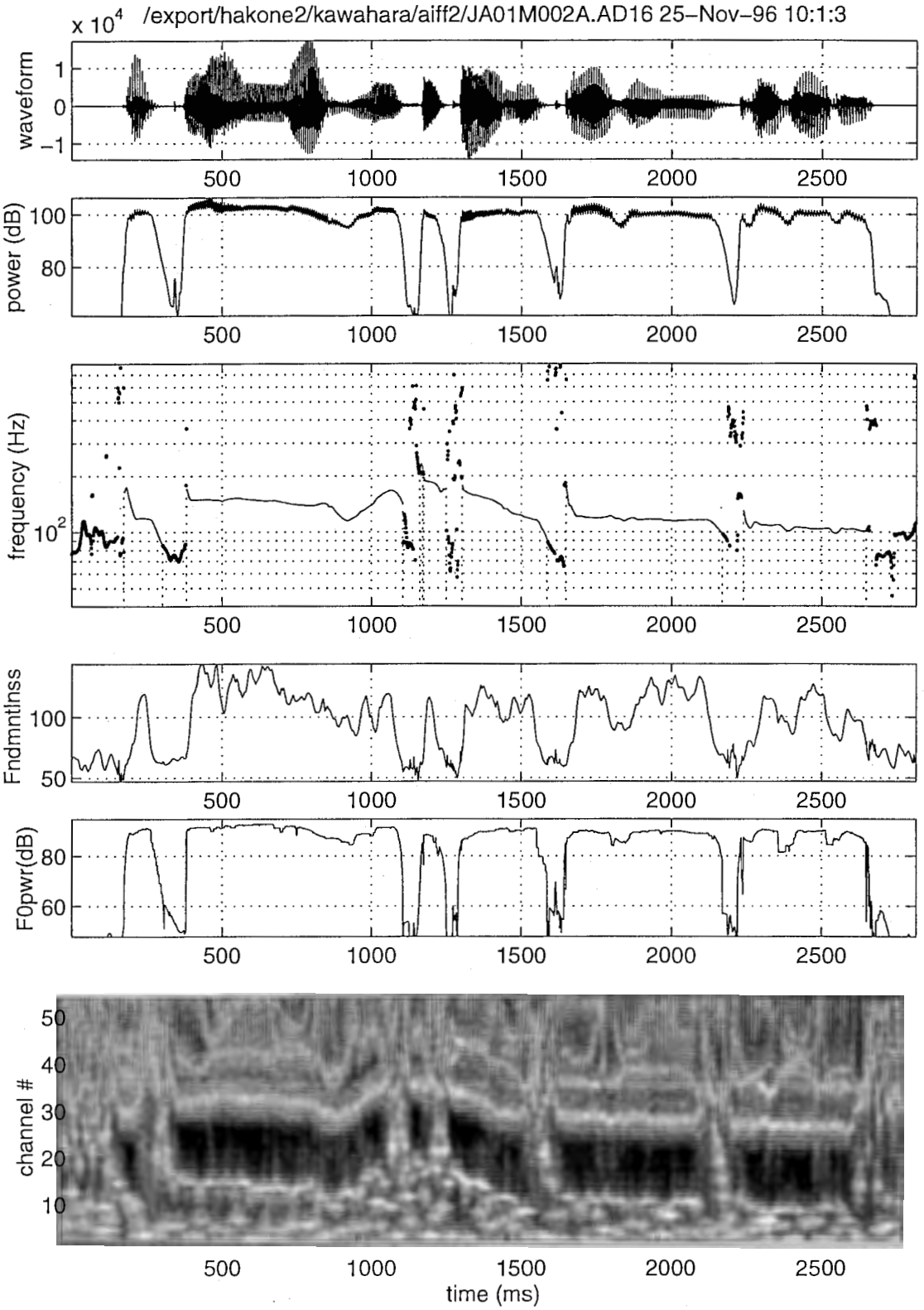

図 10: 男性の発声した音声「爆音が銀世界の高原に広がる。」の分析例。

proposed method errors			
	ordinary	subharmonic	total
NC:	2.86%	0.06%	2.92%
FHS:	0.96%	0.27%	1.23%
improved AMDF errors			
	ordinary	subharmonic	total
NC:	1.90%	0.70%	2.60%
FHS:	0.87%	1.48%	2.35%

表 2: 改良された AMDF と本方法の比較。

4.3 EGG を用いた評価

次に、ATR 音声翻訳通信研究所の Campbell らによって提供されている実際の音声と EGG が同時収録された韻律データベースの一部を用いて、英人男性 (NC) および日本人女性 (FHS) それぞれ一名が発声した各 100 文章の分析を行なった。

4.3.1 改良された AMDF との比較

共著者の一人は、簡易な計算で比較的性能が良く、神経回路としての実現ともなじみの良い AMDF に基づいて、丁寧にチューニングした方法を実現している [4]。この改良された AMDF は、上記のデータベースに対して、例えば ESPS に搭載されている既存の方法よりも優れた性能を示す。ここでは、先ず本方法と改良された AMDF との比較を行なった。

表 2に比較結果を示す。比較の対象となったのは、EGG 波形が安定した部分についてであり、NC で約 150 秒、FHS で約 240 秒分であった。EGG 波形は、0.2ms から 0.5ms の方形窓によって平滑化された後、微分、いき値処理、異常値の除去の後、声門閉止に対応するパルスが抽出され、パルス間隔が基準信号として記録された。本方法による基本周波数の値は、逆数を求めて基本周期に変換され、EGG から求められたパルス間隔と比較された。音声から求められた周期が EGG のものの整数倍の 20%以内にあるものを 'subharmonic error' として分類し、それ以外で EGG の値と 20%以上異なるものを 'gross error' として分類した。

本方法では、'subharmonic error' が非常に少ないことが特徴であり、主に倍ピッチが "ordinary error" の大半を占めている。実際 NC では基本周波数が 70Hz 以下に達する状況が頻繁に生じており、基本周波数における信号対雑音比が低下することが主な原因と考えられた。以上の結果は、後処理を全く行なわない原理そのままの実装の性能であり、この段階で既に改良された AMDF と同程度（男性）か半分（女性）の "誤り率" を達成している。

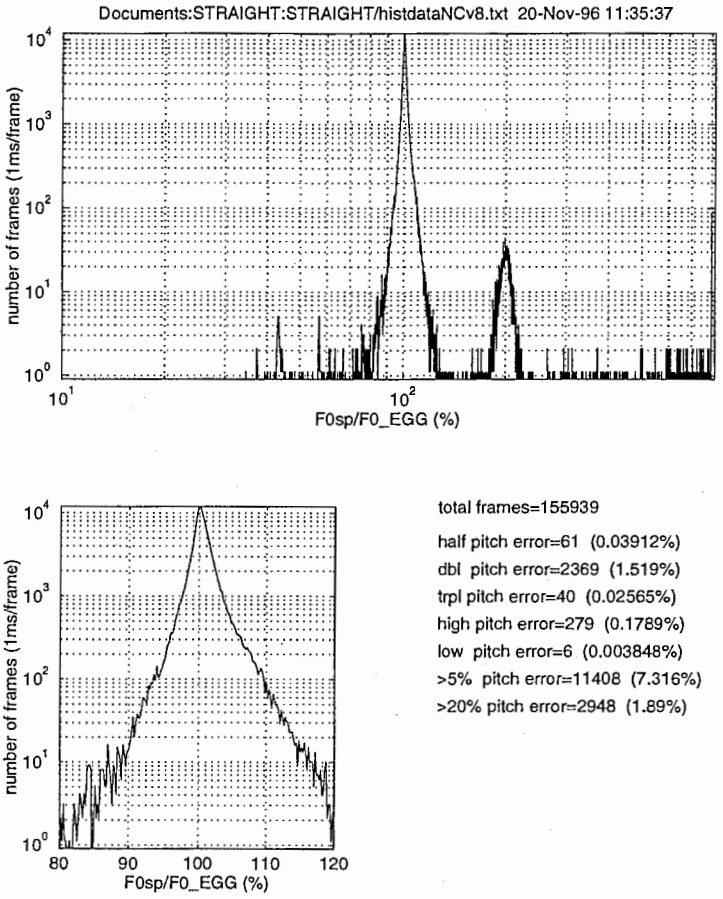

図 11: EGG からの抽出結果と音声波形からの抽出結果との比較（NC）。

4.3.2 同一の方法による EGG との比較

しかし、EGG の分析に声門の閉止に基づく方法を用いてそれを基準としてアルゴリズムを評価することは、基本周波数の定義そのものの変更を提案している本資料の立場とは矛盾する。また、閉止に基づくということは高い周波数領域の情報から基本周波数を求めていることを意味する。したがって、そのような値とそれぞれの周波数帯域毎に異なった基本周波数が存在することを前提とする本方法を用いて基本周波数成分の存在する領域の情報から求めた基本周波数とを比較することは、あまり適切な評価ではない。そこで、EGG の分析にも本方法を用い、EGG の『基本波らしさ』が 12 以上の信頼できる部分のみについて、EGG からの抽出値と音声波形からの抽出値との関連をより詳細に検討することとした。

図 11に男性話者についての分析結果を示す。図中の統計データより、基本波成分として第二調波成分が抽出される場合がかなりあることが分かる。また、低い成分よりも高い方の成分がやや抽出され易い傾向が認められる。その他の顕著な系統的な偏りの構造は見えない。この新しい基準での比較では、20%以上のずれが存在する率は 1.89%であり、改良された AMDF で用いた閉止に基づく EGG データの場合よりも良い。この新しい基準の下で EGG のデータから『基本波らしさ』が 12 以上の安定した部分として取り出されたフレー

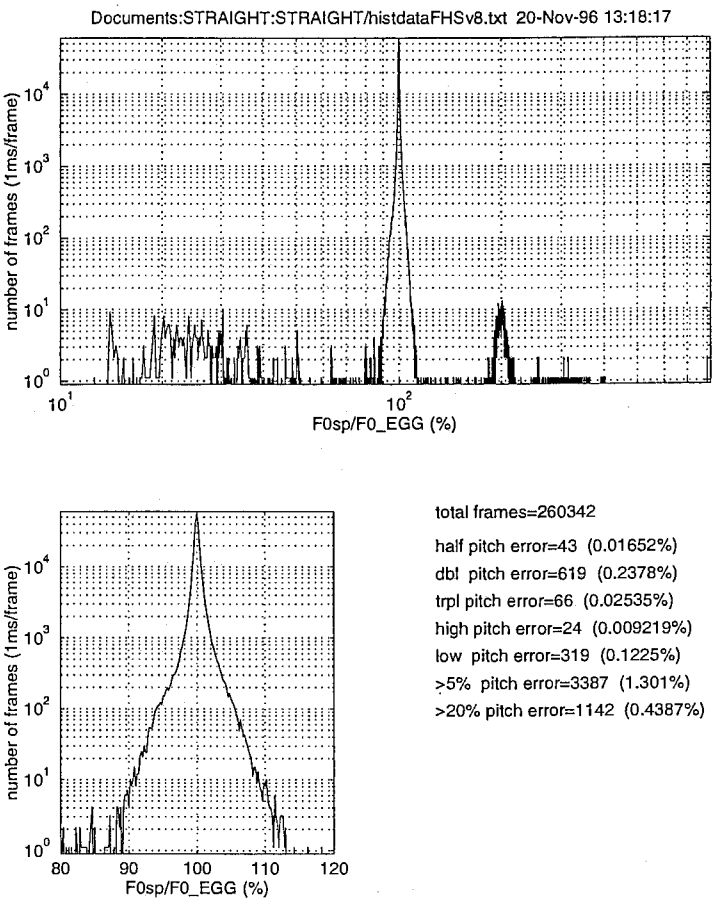

図 12: EGG からの抽出結果と音声波形からの抽出結果との比較（FHS)。

ム数は、160 秒分であり、上で説明した EGG データの解析で用いたものよりも多くなっている。すなわち、今回の評価の方がより困難な部分を多く含んでいることを意味する。

図 12に女性話者についての分析結果を示す。男性の場合と同様に、図中の統計データより、基本波成分として第二調波成分が抽出される場合がかなりあることが分かる。この場合は、低い方向への偏りが多く認められる。その他の顕著な系統的な偏りの構造は見えない。

女性の場合の偏りの少なさは、既存の方法よりもはるかに良いように見える。全分析フレームの 50%以上が EGG データと±0.3%以内の一致を示しており、±5%以上の偏りを示すのは、1.3%に過ぎない。

4.3.3 発見的知識の組み込みによる補正

そこで、これらの偏りの主要な要因である第二調波成分の抽出と高い成分を抽出する傾向を、スケール軸上でのインパルス応答を設計することにより発見的に補正することを試みた。具体的には、12 チャンネルだけ上の方に離れた部分に負のピークを有し、それ以上のチャンネルにおいて負の小さな一定値を示すようなインパルス応答関数を、発見的知識の表現とした。このインパルス応答関数を、『基本波らしさ』をある値をいき値とするなめらかな半波整流関数を通したものと畳み込んだものを補正値として用いた。

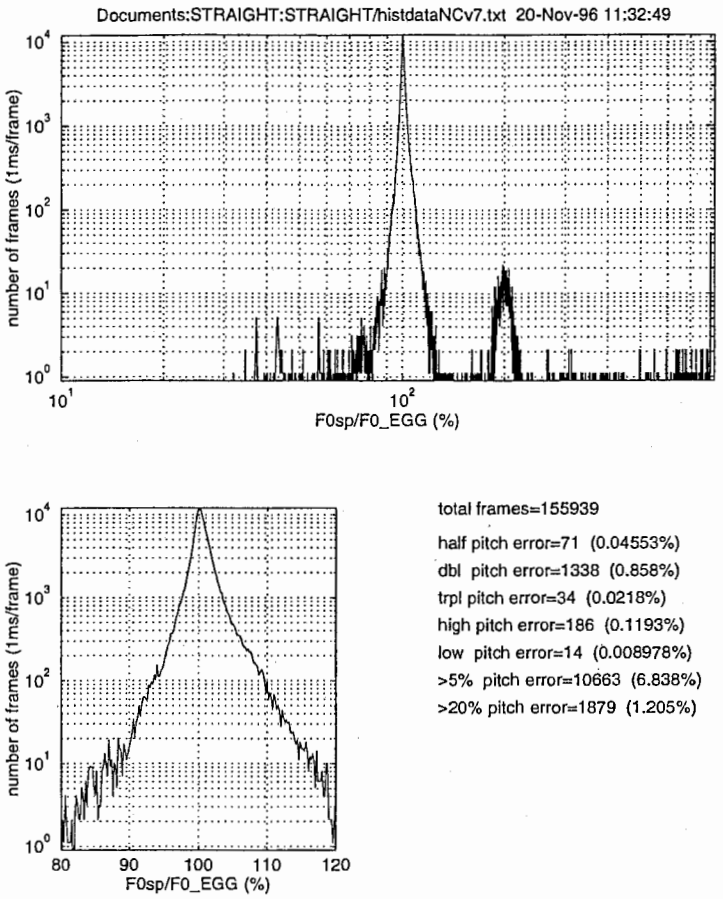

図 13: 補正後の EGG からの抽出結果と音声波形からの抽出結果との比較 (NC)。

図 13に男性話者についての補正の効果を示す。図中の統計データより、基本波成分として第二調波成分が抽出される場合が 2369 ケースから 1338 ケースへとほぼ半減していることが分かる。また、低い成分よりも高い方の成分がやや抽出され易い傾向はあまり変化していない。

図 14に女性話者についての補正の効果を示す。男性の場合と同様に、図中の統計データより、基本波成分として第二調波成分が抽出される場合が 619 ケースから 501 ケースへと 7 割程度に減少していることが分かる。しかし、補正がやや過剰であったためか、低い成分が抽出される割合が増加し過ぎ、全体としての性能は 20%以上の偏りの率で見ると、0.44%から 0.47%へとやや低下した。しかし、統計的に意味のある差ではない。また、この値自体、既存の方法と比較すると非常に良い。男性の結果と女性の結果を総合すると、20%以上の偏りは 1.17%から 0.84%に、5%以上の偏りは、4.31%から 4.08%にそれぞれ減少しており、特別な後処理無しでも、これまでの方法を十分に凌駕するか最も優れたものに匹敵する性能が実現できていることが分かる [7, 6, 16, 4]。

4.3.4 包絡成分からの基本周波数の抽出

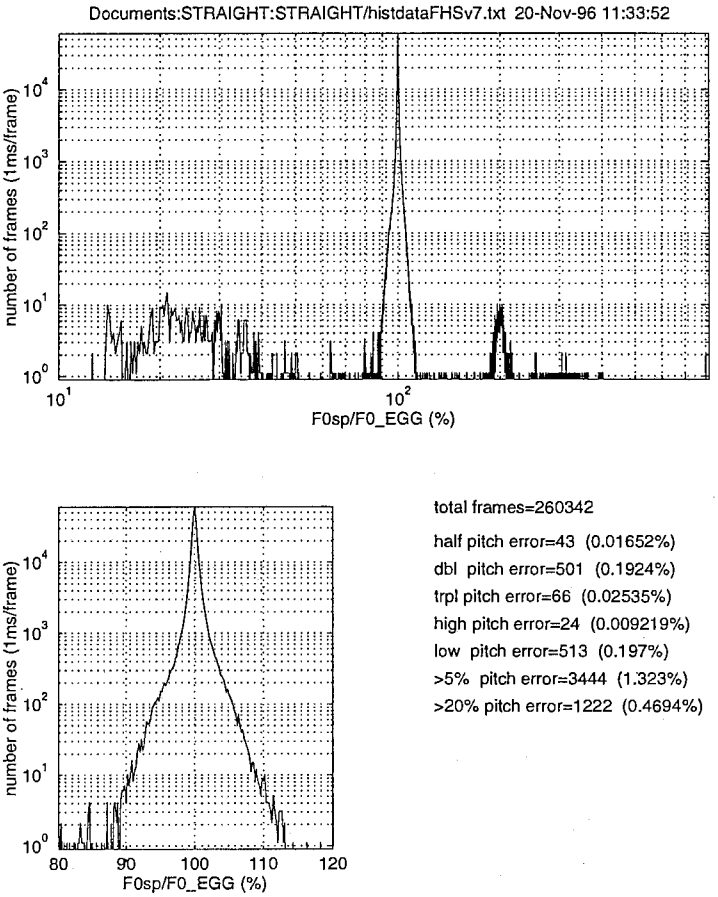

図 14: 補正後の EGG からの抽出結果と音声波形からの抽出結果との比較（FHS)。

本方法のような瞬時周波数に基づく方法は、信号の基本波成分のみから基本周波数の情報を抽出するため、他の帯域に存在する調波成分の持つ情報を利用できないという問題が指摘されていた。また、'missing fundamental' として知られる基本周波数成分を欠いた信号の場合、基本周波数を抽出できないという根本的な問題点を抱えていた。この問題は、シミュレーションで示したように、信号の包絡を信号と見なして分析することにより解決される。また、この工夫により任意の帯域の持つ基本周波数の情報を取り出すことができる。ここでは、40Hz から 800Hz の全帯域の各々のチャンネルの出力の絶対値の自乗平均値を信号波形と見なして同様な分析を行なった結果を図 15に例示する。

低い成分が多数求められている点を除けば、同様な抽出が行なわれていることが分かる。興味深い点は、発見的知識による補正を行なわなくとも第二調波成分の抽出率が半分以下になっていることである。また、左下の拡大図に見るように、平均値は正確に一致しているが、分布は広がっている。これは、前述したように、基本周波数の帯域にある基本周波数の情報と、高い周波数領域にある基本周波数の情報が異なっていることの現われであろう。これらの結果は、人間のピッチ感覚がどの様な定義に対応するものであるのか、どのような帯域の情報を主に用いているのかという興味深い問題を提起する。

Hilbert 変換を用いて包絡を計算することは論理的には最も素直ではあるが、実装上は幾つかの代替案がある。包絡は、例えば半波整流で置き換えることが可能であるし、瞬時周波

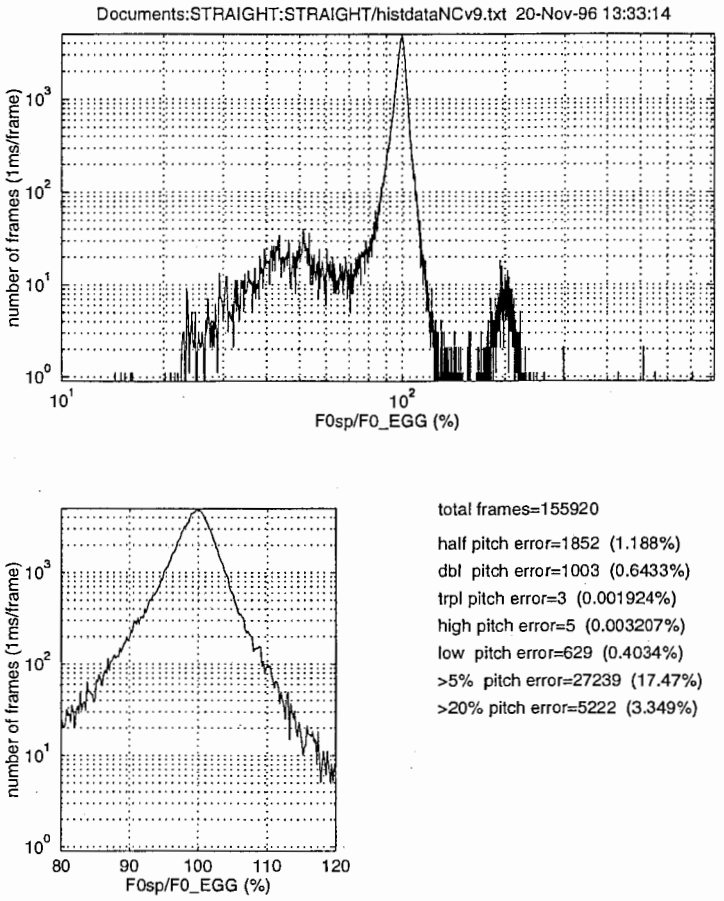

図 15: EGG からの抽出結果と音声波形の包絡成分 (40Hz～800Hz) からの抽出結果との比較 (NC)。

数や瞬時振幅 (絶対値) の計算は、ESA (エネルギー分離アルゴリズム) [13, 14] で置き換えることが可能である。これらは、計算効率が高いので、例えばハードウェアでの実装を考える場合には、有効であろう。

5　その他の要検討項目

AM と FM が最小になるような信号を基本波であると定義した。基本波が持つべき性質は、その他にも挙げることができる。最後の部分で導入した調波性は、その一つであり、分布の尖度が小さいというのも基本波の持つべき性質である。しかし、これらは AM と FM に基づく定義を本質的に拡張するものではない。実際、これらを用いた結果は、実験で述べたように常に改良方向に向かうわけではない。あくまで信号の収録条件等で生ずる雑音によるバイアスを発見的に減少させるために補助的に使用すべきものである。

5.1　聴覚の計算論との関連

本資料で提案したアルゴリズムは『基本波らしさ』に対する要請だけから出発して組み立てたものであり、聴覚についての生理学的知識や心理学的知識を用いてはいない。用いているのは、Bregman を引用した部分で、外界に存在する拘束として共通運命の法則に触れた部分である。これは、聴覚も同じ拘束を利用しているという間接的な形での関連であり、直接聴覚系についての知識を利用してはいない。しかし、もし生物が外界を聴覚を通じて理解して行く上で『基本波』というものを取り出すことが重要な意味を持っているのであれば、ここで求めた量は、結果として心理学的あるいは生理学的にも何らかの対応物を持つことになるはずである。本資料の冒頭では、瞬時周波数として求められる基本周波数と心理量であるピッチとは別のものであり、比較は不可能であると指摘した。しかし、比較は不可能であっても、それらが別のものであることを意識して慎重に議論するのであれば、対応関係を調べることは、興味深いことである。そのような対応関係がもし精密に成立するのであれば、例えば本資料での『基本波らしさ』についての議論は、聴覚でのピッチ知覚についての計算論レベルの理論を与えるものと考えて良いであろう。このような計算論の立場から逆照射することにより、聴覚の機能あるいは生理学的・心理学的な知見のいずれが本質的なものであり、いずれが本質的な機能を実現するために生体材料を利用することから生じた実装上の問題に基づく副作用であるかを明らかにできるのではないかと考えられる。

以下では、Shouten らによる非調波信号のピッチ知覚を例として、このような議論の成立の可能性を追って見る。Shouten の非調波信号は、2040Hz の正弦波に 200Hz の 100%の変調をかけたものである。その結果、1840,2040,2240Hz に成分調波が発生し、共通の F0 は、40Hz となる。しかし、知覚されるのは 204Hz 程度であると報告されている。

Schouten の信号をそのまま本方法に入力すると、キャリア周波数の 2040Hz が求められる。これは、本方法が AM-FM 信号の分析のために設計されているものなので自然な結果である。

次に、この非調波信号の Hilbert 変換による包絡を信号として本方法により分析した結果を図 16に示す。変調に用いた AM の周波数が基本周波数として抽出されている。微小な変動は、窓関数の漏れによる。基本波らしさは 30 程度と非常に高い。

包絡の計算部分に聴覚系での処理を模擬して半波整流を用い、整流された信号を本方法により分析した結果を図 17 に示す。『基本波らしさ』は、18 から 20 程度を周期的 (80Hz) で上下する。それに同期して抽出される F0 は 80Hz の周期で、198Hz から 203Hz を上下する。

同じように包絡の計算部分に聴覚系での処理を模擬した半波整流を用い、今度は 2000Hz をキャリアとする調波信号を整流した信号を本方法により分析した結果を図 17 に示す。この場合は、Hilbert 変換を用いて包絡を計算した場合と同様に、正確に AM 変調周波数とし

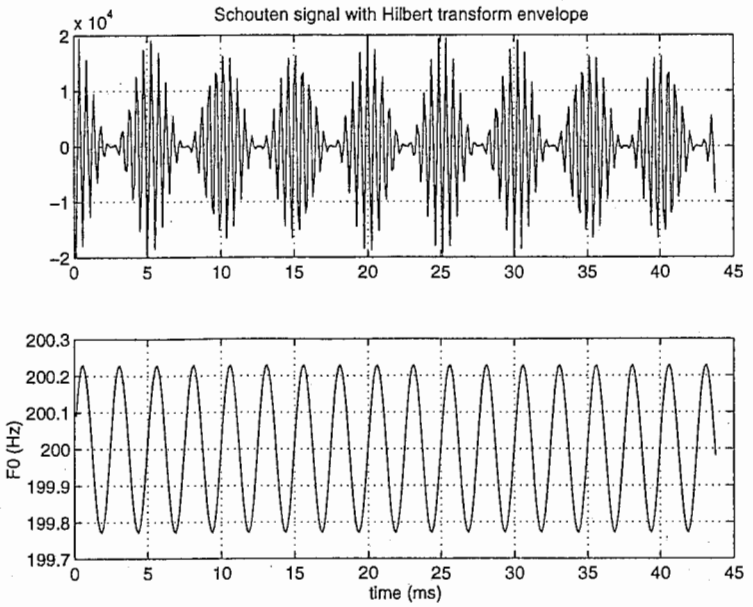

図 16: Schouten の非調波信号の Hilbert 変換による包絡信号の分析結果。AM 変調周波数が求められている。

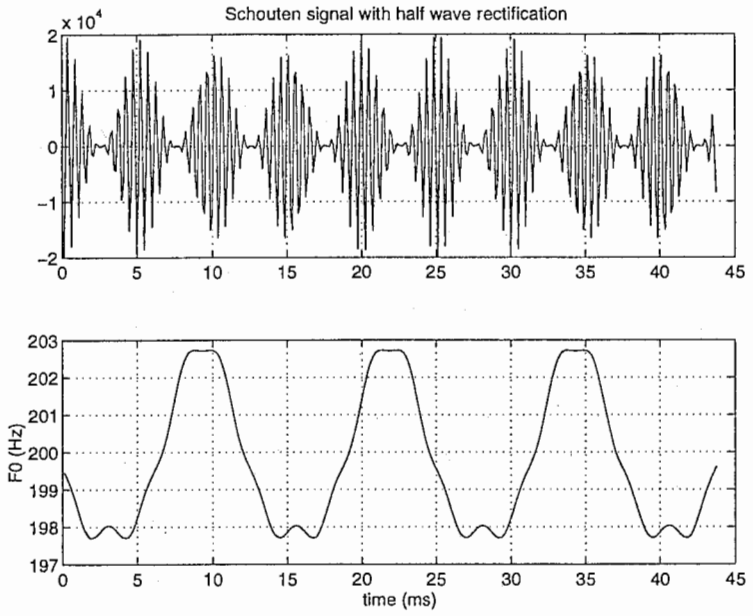

図 17: Schouten の調波信号の半波整流信号の分析結果。F0 として FM が求められている。

て用いた 200Hz が F0 として抽出される。

この一つの例だけで議論するのは危険ではあるが、以下のような仮説が成立する可能性は高いように思われる。すなわち、聴覚は、外界の音の中から『基本波らしさ』の高いものを取り出し、その瞬時周波数として抽出される現象の背後にある駆動源の基本的な速度をピッチとして知覚する。駆動源の基本的な速度の抽出には、伝播系の影響を受ける波形そ

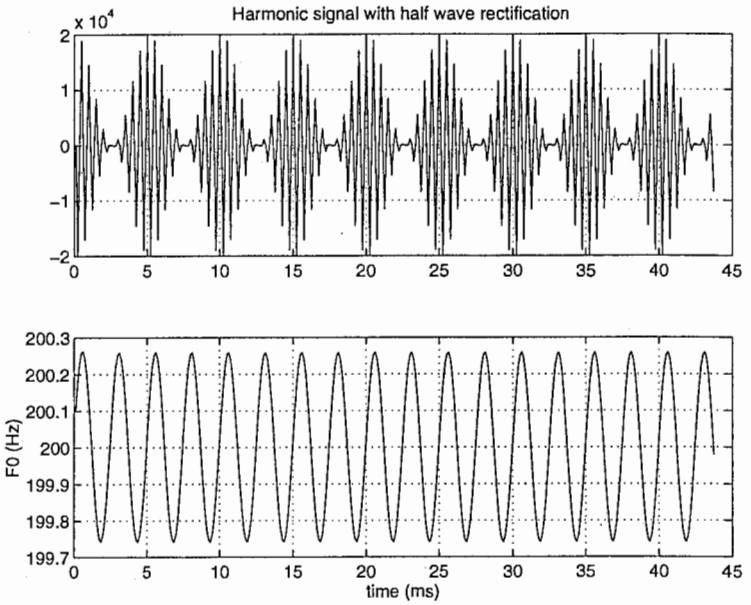

図 18: Schouten の調波信号の半波整流信号の分析結果。Hilbert 変換の包絡を用いた場合と変わらない。

のものではなく、駆動エネルギーの供給を反映する包絡情報が処理系の入力として用いられる。包絡情報の抽出には、自然界に通常存在する（疑似的に）周期的な信号では Hilbert 変換として定義される包絡と同じ結果を与える半波整流処理が、生物材料を用いた場合の実現容易性から選択されている。Schouten の非調波信号は、本来自然界では問題にならない聴覚系にとっての設計仕様外の入力であり、システムの構造をリバースエンジニアリングするためのものと理解すべきであろう。

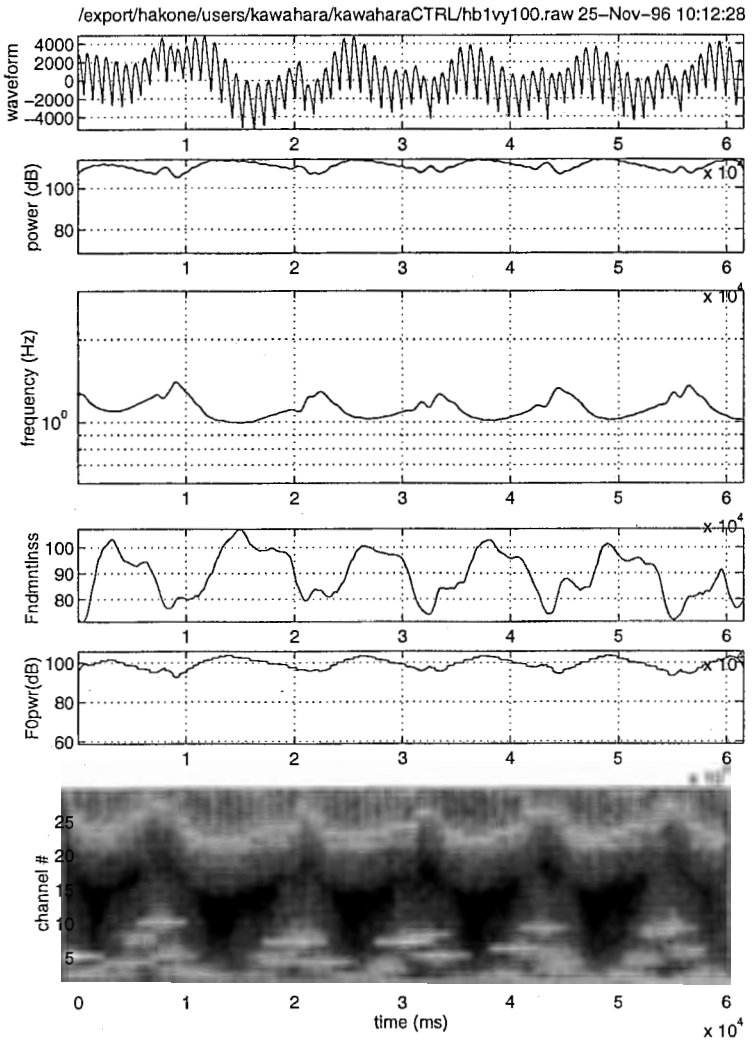

図 19: 脈波信号の解析結果。

5.2 一般の生体信号の解析

本方法は、信号の中に周期的な成分が含まれていれば定義と信号対雑音比の許す限界に近い精度でそれらの成分を抽出することができる。本方法は、音声特有の拘束条件を何も用いていないため、一般の生体信号からの周期成分の解析に適用することができる。ここでは、耳介に装着した脈波ピックアップからの信号を本方法で解析した結果を示す。記録は 60 秒間のデータで、男性被験者が母音「ア」を繰り返し発声した時に同時に収録された [11]。

心拍数は、発声に同期して周期的に変化している。図の範囲では、発声が 5 回半行なわれている。心拍数は、各発声の中央付近で最低となり、同時に最も安定した拍動状態を示していることが分かる。

6 まとめ

これまで非常に困難であると考えられてきた音声の基本周波数抽出が、解くべき問題の定義を変更することで、比較的簡単なアルゴリズム (TEMPO) により解けることを示した。今回示した方法は、実現可能性の実証のためのものであり、実用的には、計算量を2桁程度削減することが望ましい。また、このアルゴリズムを用いると、音声の異なったそれぞれの周波数領域において異なった基本周波数の抽出が可能となる。このことは、音声合成の際の音源の考え方に見直しを迫るとともに、人間のピッチ知覚に関しての興味深い問題を提起するものである。

謝辞

本資料の草稿に対して多くの有益なコメントを下さった英国 MRC の Roy Patterson 博士と、NTT 基礎研究所の入野俊夫 博士に深く感謝します。また、日頃討論して下さる ATR 音声翻訳通信研究所の片桐滋博士、渡辺秀行博士および人間情報通信研究所の多くの同僚に感謝します。

参考文献

[1] 阿部敏彦、小林隆夫、今井聖:音声信号の位相微分に基づくピッチ抽出、音講論、1-5-3、pp.237-238 (1994.10).

[2] 阿部敏彦、小林隆夫、今井聖:音声信号の非線形時間軸伸縮と瞬時周波数に基づく倍音推定、音講論、1-4-21、pp.259-260 (1995.3).

[3] Albert S. Bregman: Auditory Scene Analysis, MIT Press, (1990).

[4] Alain de Cheveigne : "Speech Fundamental Frequency Estimation", ATR Technical Report, TR-H-195, (1996).

[5] 例えば、桜井明、新井勉共訳、Charles K. Chui: An Introduction to Wavelets, Academic Press (1992). (訳書は 1993 年、東京電気大学出版局刊)

[6] Martin Cooke, Steeve Beet and Malcolm Crawford: Visual Representations of Speech Signals, John Wiley & sons, (1993).

[7] W. Hess: Pitch Determination of Speech Signals, Springer, (1983).

[8] 入野俊夫:聴覚末梢系の計算理論、聴覚研究会資料、H-95-44 (1995).

[9] 河原英紀、入野俊夫:wavelet 変換による音声の駆動情報の抽出について、音講論、3-7-8、pp.409-410 (1991.10).

[10] 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, No. EA96-28, August 1996.

[11] 河原英紀、加藤比呂子、Allan K. Barros: 音声の基本周波数揺らぎの発生源と聴覚による変形について、音講論、1-6-2、pp.375-376 (1996.9).

[12] Hideki Kawahara, Hiroko Kato and J. C. Williams: Effects of Audiotry Feedback on F0 Trajectory Generation, ICSLP'96, Philadelphia, pp.287-290 (1996.10).

[13] Petros Maragos, James F. Kaiser and Thomas F. Quatieri: "On Amplitude and Frequency Demodulation Using Energy Operators", IEEE Trans. Signal Processing, Vol.41, No.4, pp.1532-1550 (1993).

[14] Petros Maragos, James F. Kaiser and Thomas F. Quatieri: "Energy Separation in Signal Modulations with Application to Speech Analysis", IEEE Trans. Signal Processing, Vol.41, No.10, pp.3024-3051 (1993).

[15] Robert J. McAulay and Thomas F. Quatieri: "Speech Analysis/Synthesis Based on a Sinusoidal Representation", IEEE Trans. ASSP, V0l.34, No.4, pp.744-754 (1986).

[16] 都木徹、清山信正、宮坂栄一：リアルタイム音声処理のための複数窓幅による逐次ピッチ抽出法、音講論、1-1-16、pp.229-230 (1995.9).

```
%       Example pitch inspection program
%
%       If you have any questions,  mailto:kawahara@hip.atr.co.jp
%
%       Copyright (c) ATR Human Information Processing Research Labs. 1996
%       Invented and coded by Hideki Kawahara
%       30/Oct./1996
%       31/Oct./1996
%       01/Nov./1996
%       15/Nov./1996
%       19/Nov./1996 baseband conversion
%       20/Nov./1996 baseband conversion (highest two octave)
%       21/Nov./1996 baseband conversion using half wave rectification
%       25/Nov./1996 scalable version

fname=input(['Please input the speech file name: '],'s');
fs=input(['Sampling frequency in Hz:? ']);
f0shiftm=input('F0 frame rate in ms:? ');
f0floor=input('Lowest frequency to search in Hz:? ');
f0ceil=input('Highest frequency to search in Hz:? ');
dn=input('Sampling rate conversion in ratio:? ');
%f0shiftm=1;
%dn=round(fs/2000);
%dn=1;
fid=fopen(fname,'r');
x=fread(fid,'short');
fclose(fid);
%x=abs(hilbert(x));

[f01,f0dev,strand,fen,f0raw,pm]= ...
    HirbPitch9(decimate(x,dn),fs/dn,f0floor,f0ceil,f0shiftm,'off');

displaySourceInfScl

```

```
function [f0l,f0dev,strand,fen,f0raw,pm]= ...
            HirbPitch9(x,fs,f0floor,f0ceil,f0shiftm,hr);

%  [f0l,f0dev,strand,fen,f0raw,pm]=HirbPitch9(x,fs,f0floor,f0ceil,f0shiftm,hr);
%              Pitch extraction and a fundamental strand ( an auditory
%              object) extraction.
%       Input parameters
%              x       : input signal
%                        (2kHz sampling rate is sufficient. Use decimate.)
%              fs      : sampling frequency (Hz)
%              f0floor : lower bound for pitch search (60Hz suggested)
%              f0ceil  : upper bound for pitch search (800Hz suggested)
%              f0shiftm        : frame shift for F0 sampling (1ms suggested)
%              hr      : heuristics indicator 'on' or 'off'
%       Output paramters
%              f0l     : F0 information (can be used for STARIGHT)
%              f0dev   : F0 and amplitude stability indicator
%              strand  : pitch strand map. You will find auditory objects
%                                look like lakes using the following command.
%                        imagesc(strand);axis('xy');colormap(jet);
%              fen     : Energy of fundamental component (dB)
%              f0raw   : un-processed F0 information
%              pm      : wavelet analysis using iso-metric Gabor wavelet
%
%       If you have any questions,  mailto:kawahara@hip.atr.co.jp
%
%       Copyright (c) ATR Human Information Processing Research Labs. 1996
%       Invented and coded by Hideki Kawahara
%       30/Oct./1996
%       31/Oct./1996
%       19/Nov./1996 baseband conversion (highest two octave)
%       21/Nov./1996 baseband conversion using half wave rectification
%       22/Nov./1996 bug fix to make everything scalable
%       24/Nov./1996 bug fix to make everything scalable

f0=f0floor;
nvo=12;
nvc=ceil(log(f0ceil/f0floor)/log(2)*nvo);
fscale=f0*(2.0.^((1:nvc)/nvo));

t0=1/f0;
lmx=4*t0*fs; % This line is scalable
wl=2^ceil(log(lmx)/log(2));
t=((1:wl)-wl/2)/fs;


[ym,pm]=multanalytFineAG3(x,fs,f0,nvc,nvo);
[fqi,fqiv,ampiv]=instFreq3(pm,ym,fs,f0,nvo);

ampiv=diag((1/2.0).^((0:nvc-1)/nvo))*ampiv;
fqiv=diag((1/4.0).^((0:nvc-1)/nvo))*fqiv;

strand=(dB(fqiv)+dB(ampiv))-2*dB(f0);
[mm,nn]=size(strand);

MeritF=strand;
if hr(1:2)=='on'
  harmonicNB;
else
  mstrnd=-strand;
end;
[yy,mnn]=max(mstrnd);

fqis=fqi(:);
f0can=(fqis((0:nn-1)*mm+mnn));
f0can=f0can(1:fs/(1000/f0shiftm):nn);

MeritFs=MeritF(:);
f0dev=(MeritFs((0:nn-1)*mm+mnn));
f0dev=(f0dev(1:fs/(1000/f0shiftm):nn));

```



```
[pmm,pnn]=size(pm);
pms=pm(:);
pms=(pms((0:pnn-1)*pmm+mnn));
fen=(sqrt((abs(pms(1:fs/(1000/f0shiftm):nn)).^2)));

strand=(strand(:,1:fs/(1000/f0shiftm):nn));

fen=dB(fen);
f0lhb=(f0l'.*(fen>max(fen)-20));

fqired=fqi(:,1:fs/(1000/f0shiftm):nn);  % reduced fqi for later interpolation of F0
fqired=[fqired(1,:);fqired;fqired(mm,:)];  % table for interpolation

pch=min(nvc+2,max(1,(log(max(1,f0can/f0))/log(2)*nvo)+2))';

fqireds=fqired(:);
uf=fqireds((0:length(pch)-1)*(mm+2)+ceil(pch));
lf=fqireds((0:length(pch)-1)*(mm+2)+floor(pch));

f0raw=(1-abs(ceil(pch)-pch))'.*uf+(1-abs(floor(pch)-pch))'.*lf;

avfen=dBpower(sum(10.0.^(fen/10))/length(fen));

avf0=f0raw.*(fen>avfen-10);
avf0=mean(avf0(avf0>0));

f0l=f0raw.*(fen>avfen-25);

f0ji=(abs(diff(f0l))>avf0*0.05).*(1:length(f0l)-1)';
f0ji=f0ji(f0ji>0);
f0ji=[1 f0ji' length(f0raw)];
nstrm=length(f0ji);

f0l=f0raw*0;
if nstrm-1 < 1
  f0l=f0raw;
end;
for ii=1:nstrm-1
  if f0ji(ii+1)-f0ji(ii)>8
    if max(fen(f0ji(ii)+1:f0ji(ii+1)-1)) >avfen-10
      f0l(f0ji(ii)+1:f0ji(ii+1))=f0raw(f0ji(ii)+1:f0ji(ii+1));
    end;
  end;
end;

```

```
function [fqi,fqiv,ampiv]=instFreq3(pm,ym,fs,f0,nvo);

%       [fqi,fqiv,ampiv]=instFreq3(pm,fs,f0,nvo);
%               Calculate instantaneous frequencies in all channels
%               and velocity variation in frequency and amplitude
%       Input parameters
%               pm      : wavelet transform using iso-metric Gabor function
%               fs      : sampling freqnency in Hz
%               f0  : lower bound for pitch search (Hz: This have to be same as
%                               f0floor
%               nvo     : number of voices in an octave
%       Output parameters
%               fqi     : instantaneous frequency of all channels
%               fqiv    : FM power
%               ampiv   : AM power
%
%       If you have any questions,  mailto:kawahara@hip.atr.co.jp
%
%       Copyright (c) ATR Human Information Processing Research Labs. 1996
%       Invented and coded by Hideki Kawahara
%       30/Oct./1996

trmin=10/1000;   % Hypothetical minimum resolution for
                 % pitch perception (10ms This time)

trmin=0;

%ym=pm;
[mm,nn]=size(ym);
fqi=ym*0;
fqitmp=ym*0;
fqiv=ym*0;
ampiv=ym*0;
nw=round(fs/f0);

t0=1/f0;
lmx=round(4*t0*fs);
wl=2^ceil(log(lmx)/log(2));
x=x(:)';
nx=length(x);
tx=[x,zeros(1,wl)];
gent=((1:wl)-wl/2)/fs;

mpv=1;
for ii=1:mm
  t=gent*mpv;
  t=t(abs(t)<4);
  wbias=round((length(t)-1)/2);
  wd=exp(-pi*(t/max(trmin,t0*sqrt(2))).^2);

  fqi(ii,1:nn-1)=diff(unwrap(angle(ym(ii,:))))*fs/2/pi;
  tfqi(ii,1:nn-1)=diff(unwrap(angle(pm(ii,:))))*fs/2/pi;
  tmp=fftfilt(wd,abs([diff(tfqi(ii,1:nn-1)) zeros(1,wl)]*fs).^2);
  tmp2=fftfilt(wd,abs([diff(((0.0001+abs(pm(ii,1:nn-1))))) zeros(1,wl)*fs]).^2);
  tmp3=fftfilt(wd,[(((0.0001+abs(pm(ii,1:nn-2))).^2))) zeros(1,wl)]);

  tmm=[(real(pm(ii,1:nn-1))) 0.0001*randn(1,wl)];

  fqiv(ii,:)=abs(tmp(wbias+1:wbias+nn))/sum(abs(wd));
  ampiv(ii,:)=abs(tmp2(wbias+1:wbias+nn)./tmp3(wbias+1:wbias+nn));

  mpv=mpv*2^(1/nvo);
end;

%fqiv=fqiv/sum(abs(wd));
%ampiv=ampiv/sum(abs(wd));
fqiv=sqrt(fqiv);
ampiv=sqrt(ampiv);
```



```
function [ym,pm]=multanalytFineAG3(x,fs,f0floor,nvc,nvo);

%       Dual waveleta analysis using Alternating Gaussian
%               [ym,pm]=multanalytFine(x,fs,f0floor,nvc,nvo);
%       Input parameters
%                       x       : input signal (2kHz sampling rate is suffici
ent. Use decimate.)
%               fs      : sampling frequency (Hz)
%               f0floor : lower bound for pitch search (60Hz suggested)
%               nvc     : number of total voices for wavelet analysis
%               nvo     : number of voices in an octave
%       Outpur parameters
%               pm      : wavelet transform using iso-metric Gabor function
%
%       If you have any questions,  mailto:kawahara@hip.atr.co.jp
%
%       Copyright (c) ATR Human Information Processing Research Labs. 1996
%       Invented and coded by Hideki Kawahara
%       30/Oct./1996

t0=1/f0floor;
lmx=round(4*t0*fs);
wl=2^ceil(log(lmx)/log(2));
x=x(:)';
nx=length(x);
tx=[x,zeros(1,wl)];
gent=((1:wl)-wl/2)/fs;

%nvc=18;

wd=zeros(nvc,wl);
wd2=zeros(nvc,wl);
ym=zeros(nvc,nx);
pm=zeros(nvc,nx);
mpv=1;
for ii=1:nvc
  t=gent*mpv;
  t=t(abs(t)<4);
  wbias=round((length(t)-1)/2);
  wd(ii,1:length(t))=exp(-pi*((t-t0/4)/t0/1.3).^2).*exp(i*2*pi*(t-t0/4)/t0) ...
   -exp(-pi*((t+t0/4)/t0/1.3).^2).*exp(i*2*pi*(t+t0/4)/t0);
  wd2(ii,1:length(t))=exp(-pi*(t/t0).^2/2.4).*exp(i*2*pi*t/t0);
  pmtmp=fftfilt(wd(ii,:),tx);
  pmtmp2=fftfilt(wd2(ii,:),tx);
  pm(ii,:)=pmtmp(wbias+1:wbias+nx);
  ym(ii,:)=pmtmp2(wbias+1:wbias+nx);
  mpv=mpv*(2.0^(1/nvo));
end;
```

```
%       Pitch information display
%
%       If you have any questions,  mailto:kawahara@hip.atr.co.jp
%
%       Copyright (c) ATR Human Information Processing Research Labs. 1996
%       Invented and coded by Hideki Kawahara
%       30/Oct./1996
%       31/Oct./1996
%       01/Nov./1996
%       22/Nov./1996 scalable version

if ~exist('figlh')
  figlh=figure(1);
end;
figure(1);
reset(gcf);
set(figlh,'PaperPosition',[0.3 0.25 8 10.9]);
set(figlh,'Position',[30 130 520 680]);

avf0=mean(f0l(f0l>0));
trnj=(abs(diff([f0l;0]))>0.05*avf0);
trnj=trnj.*(1:length(f0l))';
trnj=trnj(trnj>0)';

subplot(811);
plot((1:length(x))/fs*1000,x,'w');grid on
axis([1 length(x)/fs*1000 min(x) max(x)]);
title([pwd '/' fname ' ' date ' ' mktstr]);
ylabel('waveform');

pwr=sum(abs(pm).^2)+0.1;

subplot(812);
plot((1:length(pwr))/fs*dn*1000,dBpower(pwr),'w');
grid on;
axis([1 length(pwr)/fs*dn*1000 max(dBpower(pwr))-45 max(dBpower(pwr))]);
ylabel('power (dB)');

bb=1:length(f0l);
bbs=1:5:length(bb);
subplot(412);
%semilogy(bb*f0shiftm,f0raw(bb),'w.',
semilogy(bb*f0shiftm,f0l(bb),'w', ...
bb*f0shiftm,(f0l(bb)==0).*f0raw(bb),'w.', ...
[trnj;trnj],[f0l(trnj)';f0l(min(length(f0l),trnj+1))']+0.1,'w:')%, ...
grid on;axis([1 length(bb)*f0shiftm f0floor f0ceil]);
ylabel('frequency (Hz)');

ncf=1;
subplot(815);plot(bb*f0shiftm,-f0dev(bb)/ncf,'w');
grid on;axis([1 length(bb)*f0shiftm min(-f0dev/ncf) max(-f0dev)/ncf])
%xlabel('time (ms)');
ylabel('Fndmntlnss ');

subplot(816);plot(bb*f0shiftm,fen(bb),'w');
grid on;
axis([1 length(bb)*f0shiftm max(fen)-45 max(fen)+2]);
%xlabel('time (ms)');
ylabel('F0pwr(dB)')

[mm,nn]=size(strand);
subplot(414);imagesc([1 length(bb)*f0shiftm],[1 mm],strand);axis('xy');
colormap(gray);
xlabel('time (ms)');
ylabel('channel #');
```



```
%       Procedure to introduce heuristics about
%       harmonic structure
%       Revised on
%       24/Nov./1996

stbias=50;   % bias for strand level
tstrnd=(-strand-stbias)/10;
nvo=12;

%-- ROEX type compression of very low level

tstrnd=(abs(tstrnd)+exp(-abs(tstrnd))+tstrnd)/2;

mstrnd=tstrnd*10;

[mm,nn]=size(mstrnd);

%dprs2=[-0.1 -0.2 -0.3 -0.2 -0.1];
dprs2=[-0.1 -0.2 -0.2 -0.2  -0.1];

dprs2=dprs2;

for ii=-1:1
  mstrnd(1+nvo+ii:mm,:)=mstrnd(1+nvo+ii:mm,:) ...
     +dprs2(ii+3)*tstrnd(1:mm-nvo-ii,:);
end;

enhs=[0.05 0.1 0.05];
for ii=-1:1
  mstrnd(1:mm-nvo-ii,:)=mstrnd(1:mm-nvo-ii,:) ...
     +0*enhs(ii+2)*tstrnd(1+nvo+ii:mm,:);
end;

for ii=nvo+6:nvo*2-1
  mstrnd(1+ii:mm,:)=mstrnd(1+ii:mm,:)-0.05*tstrnd(1:mm-ii,:);
end;

for ii=nvo*2:mm
  mstrnd(1+ii:mm,:)=mstrnd(1+ii:mm,:)-0.05*tstrnd(1:mm-ii,:);
end;
```