不動点に基づく音源情報抽出法の評価について

河原 英紀[1]，　Parham Zolfaghari[2]

[1] **和歌山大学システム工学部**/ATR/CREST, [2]CIAIR/**名古屋大学**

[1] **〒** 640-8510 **和歌山市栄谷** 930

kawahara@sys.wakayama-u.ac.jp

あらまし　　高品質な音声分析変換合成のための音源情報抽出法の検討を進めている．ここでは，昨年提案したフィルタ中心周波数からフィルタ出力の瞬時周波数への写像の不動点を利用して基本周波数を求める方法を対象として，音声波形とEGGを同時収録した話者28名，総計840文章の規模のデータベースを用いて行った評価について報告する．まず，評価の基準を明らかにするため，高品質な音声加工に適した音源情報の表現に関する議論が行われた．評価の結果，後処理をしないフレーム毎の分析の場合でも提案の方法は優れた精度と信頼性を示すこと，簡単な後処理を行うことで，大きさが20%を超える相対誤差の発生率を男性で0.54%，女性で0.32%に抑えられることが示された．次いで，相対誤差の原因についての検討が行われた．興味深い現象として，EGGによる基本周波数情報と音声波形からの基本周波数情報が子音部において系統的な乖離を示すことが見い出された．この現象は，調音に伴う声道長変動によるドップラー効果ではないかと考えられる．

キーワード　　基本周波数，基本周期，不動点，瞬時周波数，EGG

Excitation Source Information Extraction based on Fixed-Point Analyses

Hideki Kawahara[1] and Parham Zolfaghari[2]

[1]Faculty of Systems Engineering, Wakayama University/ATR/CREST, [2]CIAIR/Nagoya University

[1]930 Sakaedani, Wakayama, Wakayama, 640-8510 Japan

kawahara@sys.wakayama-u.ac.jp

Abstract　　A series of evaluations using a database that consists of EGG signals and simultaneously recorded speech signals was conducted to investigate the performance of a fundamental frequency extraction method based on a frequency domain fixed-point analysis of a mapping from carrier frequencies to instantaneous frequencies of a wavelet analysis. The EGG database consists of 14 male and 14 female talkers and has 840 sentences in total. The investigation is a step foward to refine a high-quality speech analysis/modification/synthesis system STRAIGHT. Prior to the evaluation, discussions about requirements on desiable source information representations were given as an introduction to the evaluation. The evaluation tests indicated that the baseline performance without post processing already yields a competitive accuracy and reliability and reveals that a simple post processing reduces the gross error rate down to 0.54% for male speakers and 0.32% for female speakers. An interesting hypothesis about the cause of the observed discrepancies between EGG-based fundamental frequencies and speech-based fundamental frequencies was also proposed suggesting that rapid changes in the vocal tract length due to rapid movements of articulatory organs yield the Doppler Effect.

key words　　fundamental frequency, fundamental period, fixed points, instantaneous frequency, EGG

1　はじめに

音声認識技術の応用が広がりつつある．外国語の認識では人間に匹敵する能力を示すシステムも出現している．しかし，母国語の音声知覚に関しては，未だに機械と人間との能力には本質的なギャップがあるように見える．現在の音声認識の枠組みの下での着実な研究とは別に，このような人間の優れた能力の仕組みを調べることは重要であろう．筆者らは，そのためのツールとして高品質音声分析変換合成システムSTRAIGHT[3, 6]を提案し，改良のための検討を進めている．

このようなVOCODER型のツールでは，音声信号からいかにして精度良く音源情報と伝達特性とを分離して抽出するかが鍵となる．著者らは，そのための一つの方法として，フィルタ中心周波数からフィルタ出力の瞬時周波数への写像の不動点に基づく基本周波数の抽出法を提案し[11, 5]，STRAIGHTに組み込んで試用して来た．しかし，この方法の導入は必ずしも合成音声の品質の向上に結びついていないことが報告されている[9]．現在，この問題に対して二つの側面から検討を加えている．一つは，音源情報を表現する新たな属性の提案[13, 4, 10]であり，もう一つは，現在のパラメタの利用法の最適化である．具体的には，前者は音声中のイベント抽出に基づくものであり，現状を先月の研究会で報告[10]した．今回は，後者の検討の一ステップとして，EGGデータベース[14]を用いて行った不動点に基づく基本周波数抽出法の評価について報告する．

本報告の構成は次のようになる．まず，音源情報と伝達特性をどのように分離し表現するかについて議論し，STRAIGHTでの両者の切り分け方について説明する．次いで，高品質音声加工のための基本周波数の抽出法に要請される条件について論じ，提案した瞬時周波数に基づく方法の位置付けを明らかにする．これらの背景の下，瞬時周波数に基づく方法が現在のSTRAIGHTにおいてどのように実装されているかを紹介し，幾つかの問題点について議論する．後半では，まず評価のための基準と尺度について論じた後，様々な条件における評価結果を示す．最後に，これらの結果をどのように加工音声の品質向上に結びつけるかを論ずるとともに，他の応用について議論する．なお，付録に重要な式とその説明を置いた．

2　音源情報と伝達特性の分離

音声の生成には，幾つかの異なった発音機構の音源が関わっている．また，音源から聴取位置までの伝達経路も，それぞれの音源や音の種類により異なっている．子音部分ではこれらが急速に切り替わり，母音部ではゆっくりと変化する．

音声生成器官の運動速度は，音声信号に含まれる周波数成分と比較すると十分にゆっくりとしている．したがって，もし，高い周波数成分を供給する音源信号とゆっくりと変化する伝達特性とを分離することができ，それぞれが少数のパラメタで記述できるのなら高度な情報圧縮や自由な音声の加工が容易になることが期待できる．特に音声のモーフィングや音声の感情エディタのようなものを実現しようとする時には，このような性質の良いパラメタを用いた構造的分解が必須である．

しかし，音源情報と伝達特性とを分離することは，いずれかに強い仮定を導入しなければ解くことのできない問題である．STRAIGHTでは，この問題を直接解くことを止め，音源を周期成分と非周期成分に二分して周期成分については基本周期のみで表わし，音源の大局的なスペクトル構造と伝達特性をまとめた滑らかな時間周波数表現を求めることとした．これは，聴覚が信号を周期／非周期と基本周波数という音源情報と大局的なスペクトル構造という二つの側面に分解していることを仮定し，それと同型な処理を工学的に実現したと言うこともできよう．このような割り切りが，実時間処理に向いた構成を可能にしている．

ただし，この二分法は余りにも単純である．滑らかな時間周波数表現と周期／非周期と基本周波数という二つの音源情報に加え，音源に関する第三の情報が必要なことはSTRAIGHTの初期から意識されていた[12]．実際，これらの音源情報と滑らかな時間周波数表現が精密に抽出されていたとしても，得られた基本周波数の情報をパルスの間隔として表現した信号を音源としたのでは，従来のVOCODER型の合成法と同じように，バズ的な音色が合成音声の品質を大きく損なってしまうのである．

2.1　音源の第三の属性

音源に関する第三の情報を精密に議論するには，音源のどのような属性が聴覚的に聞き分けられているかという心理学あるいは生理学の知見に立脚することが本来は望ましい．しかし，聴覚研究の問題は，そのような知見自体が信号処理技術と切り離しては獲得することができないことにある．このような状況で可能なのは，信号処理モデル作成というトップダウンの過程と，それを用いた聴覚心理実験というボトムアップの過程を連携させるブートストラップ型の研究戦略であろう．

現在，この第三の情報に関して信号処理側が持つ戦略は，マルチパルス駆動，コード駆動，正弦波モデル，MBE等である．STRAIGHTは，これに駆動パルスの群遅延特性の制御[12, 3]という手段を加えた．合成音声の駆動音源を位相制御したパルスにより作成するという提案[17]も，同じ発想に基づいたものである．マルチパルス駆動やコード駆動は，本質的には波形の再現を目指す手法であり，音源情報の聴覚的な構造化に貢献するようなものではない．正弦波モデルは，表現能力は高いものの，聴覚の理解に貢献する情報に結びつく構造を欠いている．

MBE型の分析合成システム[2]は，帯域毎に周期信号か非周期信号かを判定し，非周期成分の合成に定常雑音を用いる方法である．周期成分と非周期成分の比率を求めることで，二値的な判定を避ける方法も提案されている[7]．前報[10]で提案したような群遅延に基づく分析手法と合成時の群遅延特性の制御とをMBEのように帯域分割の下で組み合わせ，系統的な聴覚心理実験を行うことが必要な次のステップであろう．なお，現在のSTRAIGHTの実装においても，アドホックな実装ではあるが，周期成分と非周期成分の相対的な割合いを求め

て合成のための音源制御を行っている．

2.2 瞬時周波数とパルス間隔

ただし，第三の属性の有用性は，基本周波数が正確に信頼性高く抽出されてはじめて発揮されるものである．ここでは，具体的な事例の紹介を交えながら，高品質な音声加工への応用を狙った基本周波数の抽出法について説明する．

基本周波数の抽出には，様々な方法が提案されている [16]．それらは，大きく括れば，基本波の情報を利用するものと，基本周期の情報を利用するものとに分けることができる．定常的な周期信号の場合には，両者は一致する．しかし，音声のように次々と特性が変化して行くようなものの場合には，必ずしも一致しない．

基本波の情報を利用する方法では，基本波のみあるいは低次の調波成分の瞬時周波数から基本周波数が求められる．基本周期の情報を利用する方法では，白色化された自己相関関数等のピーク位置の情報から基本周期が求められる．この場合，主に高い周波数領域にある成分の持つ情報の比重が大きくなる．ここで，両者の関係を極端な例を用いて説明する．

間隔が変化して行くほぼ周期的なパルス列を考える．時間間隔に基づく方法では，分析位置の中心の前後にあるパルスの間隔が基本周期を決めるので，基本周期は階段状に変化する．基本周期は，分析位置がパルス位置を横切るときに不連続に変化する．同じ信号を基本波の瞬時周波数に基づく方法で分析すると，基本波を選択する段階で位相が平均化されるため，抽出された基本周波数は連続的に変化する．このように，両者は一見すると別の情報を抽出しているようであるが，ここで挙げた例では，同じ内容を別の形で表現しているに過ぎない．実際，瞬時周波数に基づいた調波モデルの位相を cos 位相にすると同じパルス列を復元することができる．

実音声でも同様な現象が生ずる．有声音の場合には，高い周波数領域では，声門閉止時点での呼気流の変化速度の不連続が音声のエネルギーの主要な供給源である．乱暴に近似すると，高い周波数領域の音声波形は声門閉止時点にあるパルスで駆動されているとみなすことができるのである．図1は，男性の発声した母音連鎖「アイウエオ」を時間間隔に基づく方法と瞬時周波数に基づく方法で分析した結果である．ここでは，時間間隔に基づく方法として，嵯峨山が1978年に発明した方法 [15] の変型[1]，瞬時周波数に基づく方法としては，昨年までの STRAIGHT に実装されていた方法 [6] を用いた．時間間隔に基づく方法では，基本周波数が階段状に変化するのに対し，瞬時周波数に基づく方法では，基本周波数が連続的に変化しているのが分かる．ただし，図1の下の図に見るように，瞬時周波数は階段の中点を必ずしも通ってはいない．これは，実音声の場合には，パルスの場合と異なり，時間間隔から求められる基本周波数と瞬時周波数から求められる基本周波数が本質的に同じだとは限らないことを示している．このことは，次の分析によりはっきりとする．

[1]時間間隔に基づく方法としてポピュラーな cepstrum による方法を用いても，変型自己相関関数を用いても同様な傾向は認められる．階段状の軌跡は，時間間隔を求めるという基本的な考え方に起因するものである．

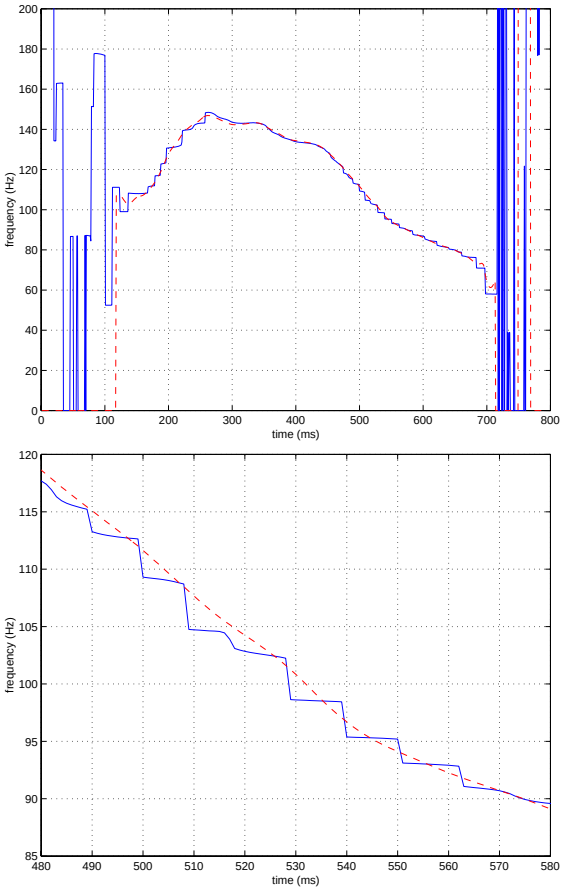

図 1: Extracted F0 trajectories for a Japanese vowel sequence /aiueo/ spoken by a male speaker using STRAIGHT (broken line) and modified Sagayama's method (solid line). -6 dB/oct equalization is used. The lower plot shows an expanded view.

図2に，音声波形と基本周波数と駆動点の位置における基本波の位相を示す[2]．端に○印のついた棒が基本波の位相を示す．所々で下向きの棒は，声門閉止時点以外の部分に対応する駆動点として求められた点に対応する．ここで，駆動点は，全報 [13] で提案した最小位相システムの群遅延による補償を用いた方法により求められた．図に見るように，声門閉止に対応する駆動点においても，対応する基本波の位相は一定ではない．これは，高い周波数領域と低い周波数領域に異なった基本周波数情報が乗っていることを示すものである．

基本周波数の情報を高品質な音声の加工に応用する場合には，滑らかなパラメタとして表現されていた方が使い易い．時間間隔に基づく方法でも，常に規則的な階段状になるのであれば，それぞれのステップの中央の位置の時刻と基本周波数を表の形で記憶させておくことで同様に使い易いパラメタとすることができる．しかし，時間間隔に基づく方法では，分析窓の大きさと基本周期の比，波形の前処理，声質等の要因により，必ずしも軌

[2]ここでは，基本周波数と位相の分布が重なっているので，左側の縦軸の目盛りを共用している．音声波形には，見易くなるようにバイアスとして 175 を加えている．

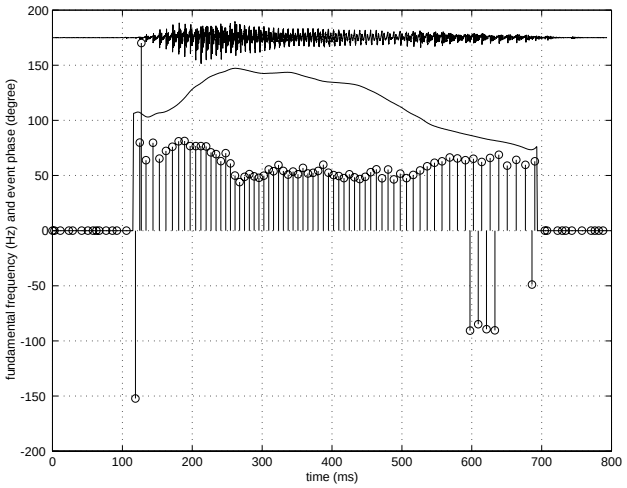

図 2: Integrated plot of the fundamental frequency trajectory, event phase and waveform.

跡が規則的な階段状になるとは限らない．そのため，補間性の良い基本周波数軌跡を時間間隔に基づく方法から求めることは一般には簡単な問題とはならない．それに対し，基本波成分の瞬時周波数から基本周波数を求める方法では，特別な後処理をせずに補間性の良い軌跡を求めることができる．この場合，基本波選択のためのフィルタの周波数選択性が鋭すぎることによる時間分解能の低下や，周波数選択性が広すぎることによる調波間の干渉が問題となる可能性がある．しかし，STRAIGHT に用いられている方法では，基本周波数の軌跡の表現に最も適した時間分解能および周波数分解能のフィルタが自動的に選択される仕組みが組み込まれているため，この問題は回避される．

3 STRAIGHT での実装

図 3に，STRAIGHT に実装されている方法（付録 A 参照）による音源情報抽出の表示例を示す [11]．試料は同じく，男性の発声した日本語母音の連鎖「アイウエオ」である．標本化周波数は 22050Hz である．不動点の抽出に用いた wavelet の個数（フィルタのチャンネル数）は 104 個であり，40Hz から 800Hz の中心周波数を $2^{1/24}$ のステップで覆っている．基本周波数は，1ms 毎に抽出されている．最上段の画面には，抽出された不動点を示す．図の横軸は時間であり，縦軸はフィルタのチャンネル番号（k）を表わしている．チャンネル番号 k とフィルタの中心周波数 λ_c は $\lambda_c = 40 \cdot 2^{(k-1)/24}$ により対応付けられる．図では白点として表示されている不動点は，基本波成分の候補の位置を表わしている．また，背景として，推定された C/N 比を濃淡画像として表示している．ここでは，雑音レベルが高い（C/N 比が小さい）ことを画像の暗さに対応させている．付録 A で述べるように，基本波成分の瞬時周波数と分析 wavelet のキャリア周波数が近い場合にだけ C/N 比が小さくなるように，分析 wavelet は設計されている．その下の段

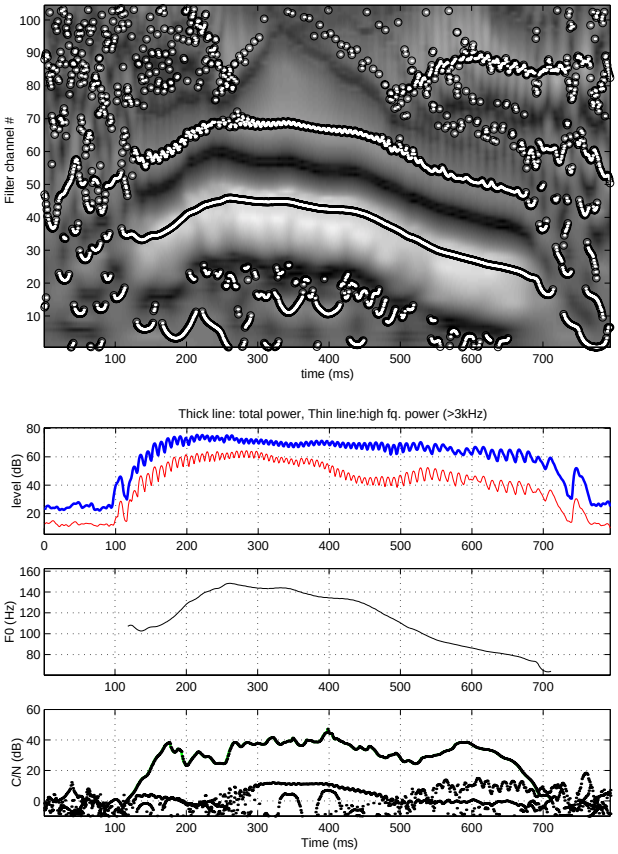

図 3: Extracted source information from a Japanese vowel sequence /aiueo/ spoken by a male speaker. The top panel represents fixed points extracted using a circle symbol with a white center dot. The overlayed image represents the C/N ratio. The lighter color indicates a higher C/N ratio. The middle panel shows the total energy (thick line) and the higher frequency ($>$ 3 kHz) energy (thin line). The next panel illustrates an extracted F0. The bottom panel shows the C/N ratio for each fixed point.

のグラフは，全体のエネルギーと 3kHz 以上の帯域のエネルギーを示し，その下の段のグラフに，抽出された基本周波数の軌跡を示している．最下段のグラフは，不動点それぞれについての C/N の推定値を示している．母音部分では基本波に対応する不動点の C/N 比だけが非常に大きな値を示し，それ以外は 0 dB 付近の値をとっていることが分かる．

上で説明したような特殊な wavelet の設計法と C/N 比の性質により，基本的には，ある時刻における不動点の中から，最も C/N 比の大きなものを選択することによって基本周波数に対応する不動点が選択される．実際，最上段の画面では，背景画像の明るい部分の上に描かれた不動点の滑らかな連なりとして基本周波数の軌跡が明瞭に認められる．1999 年版の STRAIGHT の実装の一つの問題は，帯域別のエネルギーに基づいた有声/無声のための後処理が予備実験の結果だけを用いたアドホックな実装となっていたため，視覚的に認めるこ

とのできる基本周波数の軌跡を完全には抽出できないことにあった．また，録音の条件によっては，レベルの比較的低い商用交流による電磁誘導雑音や空調雑音が混入することがある．そのような場合，雑音の周期成分の方が基本周波数成分として誤って抽出されることも多かった．特に，母音の開始や終了の部分，急速に基本周波数が変化する子音の周辺では音声の基本波成分のC/N比が急速に低下するため，このような誤りや無声側に偏った判定が生ずるという問題があった．

4 EGGデータベースによる評価

EGGと音声を同時収録したデータベース[14]を用いて，本方法の様々な実装の性能の評価を行った．データベースには，男性14名，女性14名がそれぞれ30文章を発声した合計840文章の試料が収録されている．また，EGG信号の視察による有声部分と無声部分の判定情報も時間軸を揃えて収録されている．

基準となるEGGの基本周波数データを作成するにあたっては，同時に記録されている有声／無声判定を参照し，第三調波までの瞬時周波数とC/N比を用いて推定精度の向上を図った．また，声門から唇までの伝播遅延時間の補正を行って音声から求められた基本周波数との比較を行った．評価は1ms毎に行い，男性の場合は650,284フレーム，女性の場合は643,209フレームを評価対象とした．

本方法の性能に関連するパラメタは以下のように設定した．1オクターブあたりのフィルタの個数は，予備実験の結果から6個とした[3]．waveletの作成に用いた関数は，基本周期を基準とした時間方向の相対的な広がりと，基本周波数を基準とした周波数方向の相対的な広がりが等しいような等方的なGabor関数よりもやや時間方向に伸長した関数から作成した．時間方向の伸長率ηは，1.2とした．

ここでは，後処理を含め，4種類の実装について評価を行った．それらは，(1) 後処理無し(O)，(2) 不動点の連続性に基づく探索とC/Nに基づく選択(C)，(3) (2)に第三調波までの情報に基づく基本周波数の修正を加えたもの(R)，(4) STRAIGHTでの実装(S)，である．

4.1 抽出された基本周波数の分布

図4に，音声波形から求められた基本周波数をEGG波形から求められた基本周波数で正規化した値の分布を示す．上の図は男性，下の図は女性についての結果である．図中の実線は基本波から第三次調波までを用いて基本周波数を求めた場合（R），破線は基本波のみから基本周波数を求めた場合（C）の分布を示す．後処理をしない場合の分布は，ほぼ（C）と同じで，画面から外れるデータの割合が増加していることが違うだけである．これらの結果によれば，複数の調波の利用による推定値の改善効果は，女性の分布の裾の部分に見られるのみであり，それほど大きくはない．これは，今回の評価に用いた資料の音声の収録条件が良く，含まれる雑音のレベルが低かったためであろう．なお，たて軸を対数表示としているため，誤差の大きな部分が目立つが，実際には，分析対象の80%以上のフレームで相対誤差の大きさが1%以下であるという，非常に鋭い分布である．

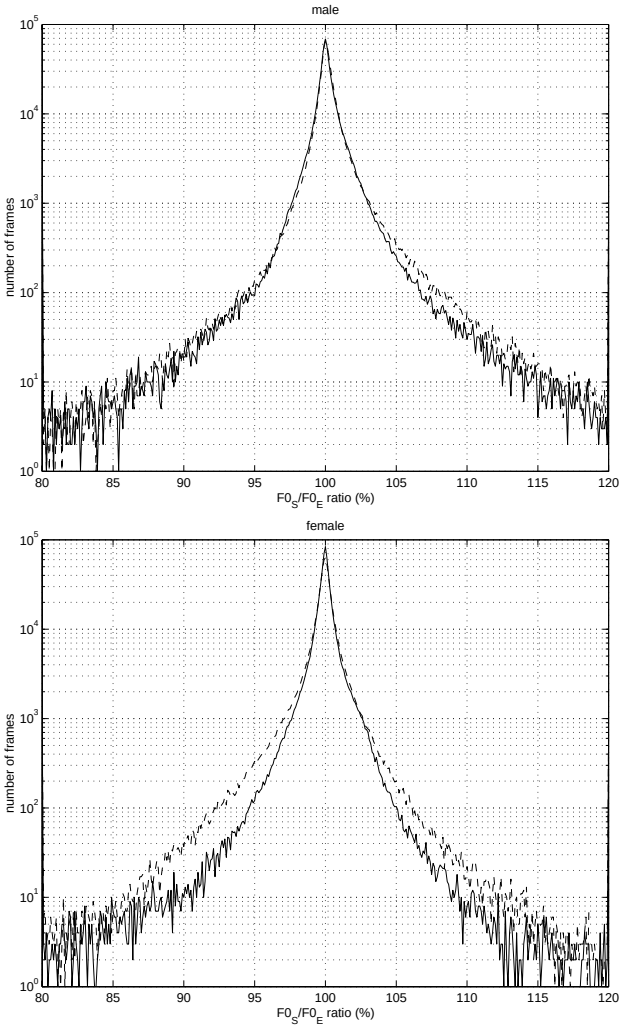

図 4: Distribution of extracted F0 normalized by the reference F0 basede on EGG data. Solid line represents the distribution after F0 refinements using lower three harmonic components. Dashed line represents the distribution using only the fundamental component. Upper plot shows results for male and the lower plot shows results for female.

4.2 Gross errorの話者依存性

基本周波数の推定の大きな誤りは，分析合成型の音声処理における主要な品質劣化の原因となる．ここでは，音声波形から求められた基本周波数がEGGから求めた値から20%以上離れた場合をgross errorとして計上する．音源情報の抽出精度は話者による違いが大きいので，ここでは，話者別に集計した結果を示す．

図5にgross errorの分析結果の例を示す．上の図が男

[3]STRAIGHTではオクターブあたり24個のフィルタを用いている．予備実験の結果はフィルタを6個としても顕著な性能劣化が無いことを示した．現在のSTRAIGHTでの実装は，情報の可視化のためには有用でもやや過剰品質である．

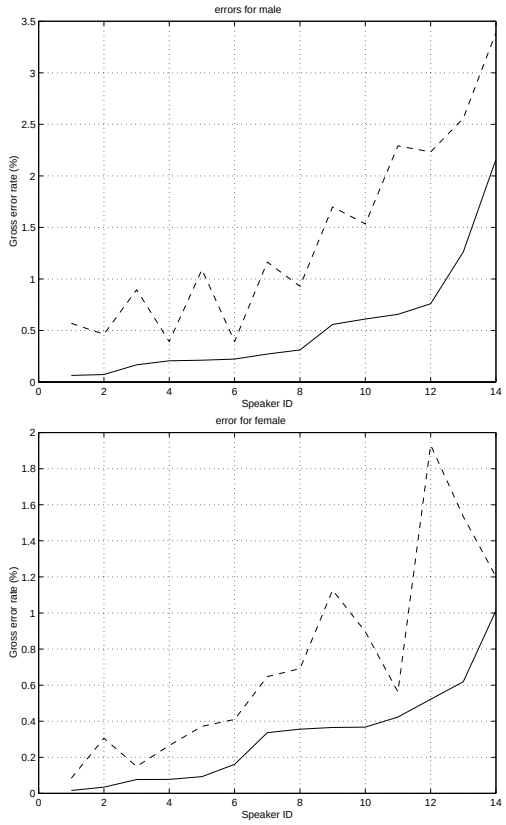

図 5: Individual gross F0 estimation error rate. Dashed line represents sores without post processing. Solid line repesents scores with fixed-point reselection and F0 refinement. Upper plot shows results for male and the lower plot shows results for female.

性の結果，下の図が女性の結果を示す．図中の破線は，何も後処理を行わない場合（O），実線は，不動点の連続性に基づいた後処理を行い，第三調波成分までを用いて推定値を改良した場合（R）の結果である．gross error に関しては，この複数調波成分の利用の効果はほとんど無い．gross error の平均値は，処理無し（O）の場合，男性で 1.40%, 女性で 0.72%である．後処理を加えた場合（R）には，これらの値は，男性で 0.54%, 女性で 0.32%となる．なお，1999 年版の STRAIGHT に実装しているアルゴリズムによる結果（S）は，処理無し（O）の場合よりも悪いものとなった．必要に迫られて設定しておいた有声/無声の判定論理が，少数の事例に基づいたものであったため，偏ったものであったようである．

4.3 Gross error の事例

gross error において最大の誤り率を示した話者においては，声帯振動そのものが異常な挙動を示している場合が幾つか見つかっている．具体例を図 6に示す．ここでは，二回分の声帯振動が組となっており，短い周期と長い周期が交番しているため，音声波形からの抽出に失敗してしまっていることが分かる．このような異常な

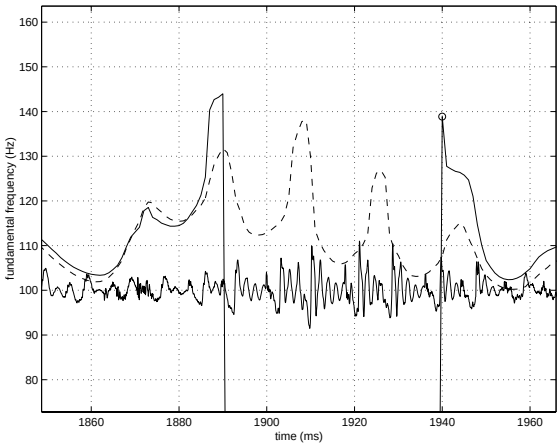

図 6: Abnormal phonation and resulted F0 error found in a male speech example. Upper lines show extracted F0 from the EGG signal (dashed line) and the speech signal (solid line). The speech signal is also shown as a solid line that centered around 100 Hz.

振動を多く含む話者を除くと，gross error の平均値は，男性で 0.4%, 女性で 0.2%となる．このような異常な場合を除いた gross error は，主に有声部分の終了部分に集中している．それらの部分では，音声の振幅も減少しており，エラーによる品質への影響は大きくはない．

4.4 EGG による F0 と音声からの F0 の乖離

Gross error が発生していない部分においても，これまで見て来たように，EGG から求められた基本周波数と音声波形から求められた基本周波数は必ずしも一致しない．この一因は，基本波成分という同じ帯域にある成分であっても，EGG 波形と音声波形では異なった情報を見ていることにある．実際，EGG が声門の閉じている区間での声門の接触面積の変動の情報を有しているのに対し，音声波形は，声門の開いている区間での声門開口部分の面積の変動の情報を有している．特に，発声の終了部分等，基本周波数と波形の振幅が同時に変化する場合には，両者の相関に依存して求められる瞬時周波数が変化し，系統的な推定誤差が生ずる．

もう一つの要因は，声道長の変動にある．顕著な例は，母音と鼻子音の接続部分で生ずる．声門から開口部分までの距離は，母音発声時と鼻子音発声時では，後者の方が約 4cm 程度長い [1]．この声道長の変化が声門の一周期の間に生ずると，例えば男性の場合，約 2%のドップラーシフトが生ずる．図 7に示す実際の分析例では，5%のドップラーシフトが生じている．1470 ms 付近の母音から鼻子音への遷移では経路長が増加するために下降方向のシフトが，1540 ms 付近での鼻子音から母音への遷移では経路長が減少するために上昇方向のシフトが生じている．

同様なドップラー効果は，図 8に示す/r/の遷移の部分でも生じている．この場合も，舌先の口蓋への接触に伴う声道形状の急速な変化による効果と考えられる．ただし，これらの現象がドップラー効果であることを検証

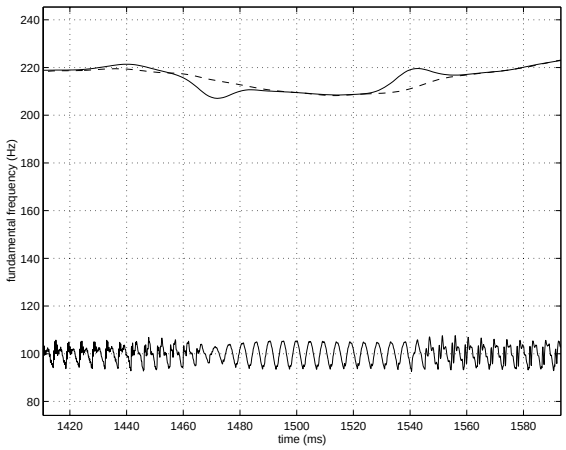

図 7: Doppler effect on F0 estimation. The plot shows a VCV transition of /ema/ excerpted from a Japanese sentence spoken by a female speaker. Upper lines show extracted F0 from the EGG signal (dashed line) and the speech signal (solid line). The speech signal is also shown at the bottom of the same plot.

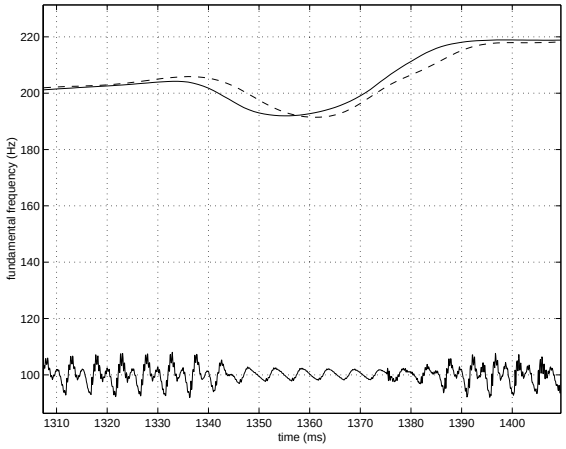

図 8: Doppler effect on F0 estimation. The plot shows a VCV transition of /ire/ excerpted from a Japanese sentence spoken by a female speaker. Upper lines show extracted F0 from the EGG signal (dashed line) and the speech signal (solid line). The speech signal is also shown at the bottom of the same plot.

するためには，同時に生ずる振幅周波数特性の変動から計算される最小位相システムの群遅延の変動による影響を補償してもこの現象が残ることを示す必要がある．

EGGから求められた基本周波数と音声波形から求められた基本周波数には，これらに起因する系統的な乖離が生ずるため，数%以下の範囲の誤差を議論することは余り意味がない．gross errorがある水準を満たしてさえいれば，むしろ系統的な偏りや，時系列としての滑らかさ等，応用側の要求条件から評価すべきであろう．

5　高品質音声処理と音源情報の表現

以上，不動点に基づく基本周波数の抽出法の評価結果を紹介した．後処理を含めた本方法は，少なくとも高品質な録音音声については，実用上十分な精度と信頼性を持つ基本周波数の抽出法であると言えるであろう．ここで紹介した後処理そのものは，2段階の処理であるため，そのまま実時間アルゴリズムとすることはできない．しかし，予備的な検討によれば，約30msの先読みを用いることで，実時間型のアルゴリズムによっても同程度の性能を実現できそうである．

STRAIGHTのパラメタと品質との関連を分析するためにDRTを用いた評価実験が行われた[8]．評価対象となったSTRAIGHTは，1998年に実装が凍結された版(v27r1)である．実験結果は，まず無声破裂音の破裂やVOTの再現に弱点があることを示し，また，無声摩擦音の周波数分解能が不足していることを示唆した．これらの問題への対策として，(1) より精度の高い基本周波数抽出法の導入，(2) 破裂子音用の適応的非対称時間窓の導入，(3) 非周期成分分析用の時間窓の寸法の拡大，(4) 周期成分と非周期成分の混合音源モデルの導入，などを行い，1999年版(v30kr16)が作成された．しかし，DRTで指摘された明瞭度に関する問題点は解消されたものの，総合的な品質の評価試験では，従来の幾つかの版と現行の版（ただし，(4) を外したもの）の間での有為な改善は認められないという結果であった．

今回の検討により，1999年版での実装では不動点に基づく方法の本来の優れた性能が生かされていないことが明らかとなった．今後は，本報告での検討結果と，イベントに基づく音源の特徴付けを組み合わせることで，聴覚的に意味のあるオブジェクトを特定し，それらの構造化とパラメタの推定法の確立を通じて総合的な品質の改善を図って行く予定である．これは，少なくとも『音声』という一つのカテゴリに関して聴覚の情景分析を実装すること，言い換えれば「聴覚脳」を実現することに他ならない．

6　まとめ

現在のSTRAIGHTに実装されている周波数領域での写像の不動点に基づくF0抽出方法の評価について報告した．その結果，実装の際の後処理が性能の劣化を招いていること，未処理あるいは簡単な後処理後の性能は非常に高く，最終的な誤抽出率を0.2 %から0.4 % とすることができることが示された．安定にこの程度の性能が実現できれば，音声応用における基本周波数情報の応用領域が大きく拡大するものと考えられる．

謝辞：本研究は，CRESTの「聴覚脳プロジェクト」の一環であり，また，一部に科学研究費(基盤C:11650425)の支援を受けた．草稿をチェックし議論してくれた山本英里博士に感謝します．

参考文献

[1] Jianu Dang, Kiyoshi Honda, and Hisayoshi Suzuki. Morphological and acoustical analysis of the nasal and

paranasal cavities. *J. Acoust. Soc. Am.*, Vol. 96, pp. 2088–2100, 1994.

[2] Daniel W. Griffin and Jae S. Lim. Multiband excitation vocoder. *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 36, No. 8, pp. 1223–1235, 1988.

[3] 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.

[4] Hideki Kawahara, Yoshinori Atake, and Parham Zolfaghari. Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay. In *Proc. ICSLP'2000*, Beijin, 2000. [to appear].

[5] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proc. Eurospeech'99*, Vol. 6, pp. 2781–2784, 1999.

[6] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. 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.

[7] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 1, pp. 1–11, January 1998.

[8] Parham Zolfaghari, 河原英紀. Subjective evaluation of STRAIGHT. 音響学会秋季講演論文集, 第I巻, pp. 193–194, 1999.

[9] 阿竹義徳, 陸金林, 中村哲, 鹿野清宏, 河原英紀. STRAIGHTの分析合成方式パラメタの主観評価による検討. 音響学会春季講演論文集, 第I巻, pp. 205–206, 2000.

[10] 河原英紀, Parham Zolfaghari. 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, 8 2000.

[11] 河原英紀, Parham Zolfaghari, Alain de Cheveigné, Roy D. Patterson. 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, 7 1999.

[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. 聴覚研究会資料, H-96-79, pp. 1–8, 1996.

[13] 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, 1999.

[14] 阿竹義徳, 入野俊夫, 河原英紀, 陸金林, 中村哲, 鹿野清宏. 調波成分の瞬時周波数を用いたピッチ推定方法の検討. 信学技報, SP99-170, 3 2000.

[15] 嵯峨山茂樹, 古井貞熙. ラグ窓を用いたピッチ抽出の一方法. 信学全大, 5, p. 263, 1978.

[16] 鈴木久喜. ピッチ抽出の今昔. 日本音響学会誌, Vol. 56, No. 2, pp. 121–128, 2000.

[17] 濱上知樹. 音源波形形状を高調波位相により制御する音声合成方式. 日本音響学会誌, Vol. 54, No. 9, pp. 623–631, 1998.

A 不動点に基づくF0抽出法

ここでは，周波数領域の不動点に基づいて基本周波数を抽出する方法についての前報[11, 5]の内容を要約して説明する．

中心周波数がλである任意の帯域通過フィルタを考える．もしフィルタの通過帯域内に周波数がλ_oであるような顕著な正弦波成分が存在すれば，フィルタ出力の瞬時周波数ωは，その正弦波成分に支配されて，ほぼその正弦波成分の周波数となる．したがって，フィルタの中心周波数λからフィルタ出力の瞬時周波数への写像$\omega(t,\lambda)$を考えると，正弦波成分の瞬時周波数の集合$\Lambda(t)$は，この写像の次のような不動点として求められることが分かる．

$$\Lambda(t) = \{\lambda \mid \omega(t,\lambda)=\lambda,\ \frac{\partial\omega(t,\lambda)}{\partial\lambda}<1\}\ , \quad (1)$$

ここで，キャリア周波数λ_cに関して等方的（あるいはやや時間軸方向にη倍だけ伸長した）Gabor関数$w(t,\lambda_c)$と，その周波数の逆数の2倍の寸法の2次のカーディナルスプライン関数$h(t,\lambda_c)$を畳込んで作成した次のような関数$w_s(t,\lambda_c)$を用いてwavelet分析を行うものとする．

$$w_s(t,\lambda_c) = w(t,\lambda_c) * h(t,\lambda_c)\ , \quad (2)$$

$$w(t,\lambda_c) = e^{-\frac{\lambda_c^2 t^2}{4\pi\eta^2}} e^{j\lambda_c t}\ ,$$

$$h(t,\lambda_c) = \max\left\{0, 1-\left|\frac{\lambda_c t}{2\pi\eta}\right|\right\}\ , \quad (3)$$

すると，このようなwaveletの構成法から自ずと，キャリア周波数が基本周波数にほぼ一致するスケールのwavelet以外では，調波を構成する複数の正弦波成分間の干渉，あるいは背景雑音レベルの相対的上昇により，主要な正弦波成分とそれ以外の成分のレベル比として定義されるC/N比（carrier to noise ratio）が増加するという性質が生ずることが分かる．

ところで，このC/N比は，不動点の周波数をλ_oとしたとき，不動点近傍での写像の幾何学的性質に基づいて以下のようにして求められる$\bar{\sigma}$を用いて$C/N = 1/\bar{\sigma}$として近似的に推定することができる．

$$\bar{\sigma}^2(t,\lambda) = \int_{-T_w}^{T_w} |w(\tau,\lambda)|\tilde{\sigma}^2(t-\tau,\lambda)d\tau \quad (4)$$

$$\tilde{\sigma}^2(t) = c_a\left(\frac{\partial\omega(t,\lambda)}{\partial\lambda}\right)^2 + c_b\left(\frac{\partial^2\omega(t,\lambda)}{\partial t\partial\lambda}\right)^2 \quad (5)$$

$$c_a = \frac{1}{\int_{-\infty}^{\infty}\left(\lambda_o \left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=\lambda_o}\right)^2 d\lambda_o}\ .$$

$$c_b = \frac{1}{\int_{-\infty}^{\infty}\left(\lambda_o^2 \left.\frac{dg(\lambda)}{d\lambda}\right|_{\lambda=\lambda_o}\right)^2 d\lambda_o}\ .$$

ここでT_wは，関数$|w(\tau,\lambda)|$が実質的に0でない範囲を覆うことができるように設定する．

これらを利用することで，基本周波数を安定に抽出することができるというのが提案した方法の基本的原理である[5]．なお，複数の調波成分から基本周波数の情報が得られる場合には，それぞれの成分についてのC/Nを用いて加重平均することで，基本周波数の推定値に含まれる誤差を減少させることができる．