社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

占有度を用いた耐雑音性の高い基本周波数推定法

中谷　智広†　　入野　俊夫†,††

† 日本電信電話 (株) NTT コミュニケーション科学基礎研究所
〒 619-0237 京都府相楽郡精華町光台 2-4
†† CREST

E-mail: †{nak,irino}@cslab.kecl.ntt.co.jp

あらまし　本稿では、背景雑音に加えてスペクトル変形を伴なった入力音声に対しても、頑健かつ精度良く基本周波数 (F_0) を推定するための新しい方法を提案する。このため、各調波成分が近傍の周波数帯域において背景雑音の影響を受けていない度合いを示す尺度である占有度 (*degree of dominance*) を、瞬時周波数に基づき定義する。占有度を用いることで信頼できる調波成分を容易に選択できるように なり、これに基づき頑健に F_0 推定を行うことができる。評価実験では、白色雑音下またはマルチトーカ雑音下での入力音に、電話音声を模擬する SRAEN フィルタによるスペクトル変形を与えた場合と与えない場合について、F_0 正解率、 および F_0 の実効誤差の評価を行った。実験結果より、提案法は、あらゆる条件下において、従来法と比べて良い結果が得られることを示す。

キーワード　占有度、基本周波数推定、瞬時周波数、背景雑音、スペクトル変形、マルチトーカ雑音、SRAEN フィルタ

Fundamental Frequency Estimation Based on Dominance Spectrum

Tomohiro NAKATANI† and Toshio IRINO†,††

† NTT Communication Science Labs., NTT Corporation,
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan
†† CREST

E-mail: †{nak,irino}@cslab.kecl.ntt.co.jp

Abstract　This paper presents a new method for robust and accurate fundamental frequency (F_0) estimation in the presence of background noise and spectral distortion. For this purpose, *degree of dominance* and a *dominance spectrum* are defined based on instantaneous frequencies of the STFT spectra. The degree of dominance is a measure for evaluating the magnitude of individual harmonic components relative to the background noise. The fundamental frequency is correctly estimated from reliable harmonic components easily selected in the dominance spectra. Experiments are performed using white and multi-talker background noise under the conditions with and without spectral distortion produced by a SRAEN filter. Results show that the present method is better than the commonly-used conventional methods in terms of both the F_0 correct rates and fine F_0 errors.

Key words　degree of dominance, fundamental frequency estimation, instantaneous frequency, background noise, spectral distortion, multi-talker noise, SRAEN filter

1. はじめに

実環境における高精度な基本周波数 (F_0) 推定は音声信号処理において重要な課題の一つである。例えば、近年、提案されている高品質ボコーダ STRAIGHT において、高精度な F_0 推定は理想的なスペクトル推定に不可欠である [1]。また、F_0 は、実環境中の混ざった音から、目的音を取り出すための重要な手がかりとされており [2]、実際に、我々は、F_0 に基づく音源分離システムを前処理として用いることで、音声認識システムの認識率が向上することを示した [3]。この音源分離においても、F_0 推定誤差が分離音の品質に与える影響は極めて大きく、混ざった音から正確な F_0 推定は重要な課題である。

これまで多くの F_0 推定法が提案されている [4]〜[7]。大別して、時間領域法と周波数領域法の 2 つに分けられる。前者は、主に入力信号の自己相関に基づく方法で、ML 法などがある。後者は、主に周波数ピークの抽出に基づく方法で、ケプストラム法などがある。どの方法が最適であるかについては現在も議論が分かれており、なおも、新しい推定法が提案され続けている。

F_0 推定性能を劣化させる音声信号の変形要因は、主に、背景雑音のような加法的雑音と空間音響や電話伝送特性のような乗法歪みが考えられる。実環境では、この両方に対して頑健性の高い F_0 推定法が望まれる。しかし、従来手法では、これら両方の種類の歪みに対して耐久性の評価がきちんと行われていなかった。また、信頼性の高い F_0 の正解値が得られる大規模な音声データベースがなかったことも問題であった。更に、多くの耐雑音性の評価では、実際の背景雑音とは性質が大きく異なる白色雑音のみが用いられており、実環境での性能が必ずしも評価できていなかった。こうした評価法の限界が、最良の F_0 推定法を決定する障害になっていたといえる。

本稿では、音声と同時に録音した electro glottal graph (EGG) 信号からなる大規模データベースを用いて F_0 推定法を評価する。耐雑音性の評価については、白色雑音に加えて、カクテルパーティを模擬するマルチトーカ雑音を用いる。さらに、音声信号に対するスペクトル変形の影響の評価のために、SRAEN フィルタを適用する。SRAEN フィルタは電話マイクの特性を模擬する 300 Hz 以上の高域通過フィルタである [11]。これらの条件を組み合わせることで、各 F_0 推定法の頑健性をあらゆる角度から比較評価する。

1.1 瞬時周波数を用いた F_0 推定法

本稿で新たに提案する F_0 推定法は、瞬時周波数 (Instantaneous frequency, IF) を用いた周波数領域法の一種である。IF とは、信号を $s(t) = a(t)e^{j(\phi(t)+\theta)}$ と表したときに、$\dot{\phi} = \mathrm{d}\phi/\mathrm{d}t$ と定義される。IF を用いた F_0 推定は、Charpentier 等によってはじめて考案され [8]、近年になって、特に、雑音下での有効性について報告されるようになった [1], [9], [10]。これらの方法において、IF は、各入力フレームごとの短時間フーリエ変換 (Short-Time Fourier Transformation, STFT) における周波数 bin ごとに計算される [13]。

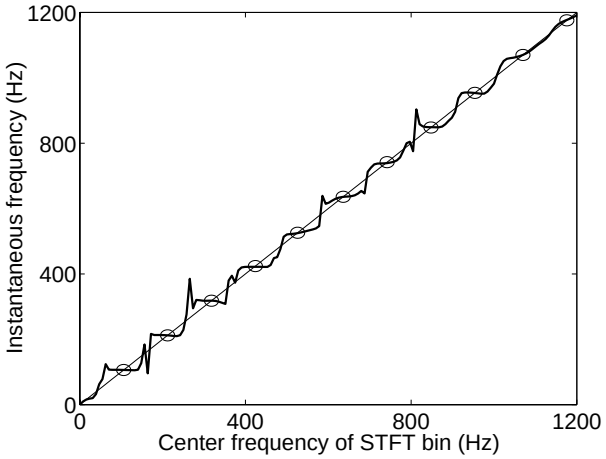

図 1 有声音の瞬時周波数 (太線) と不動点 (円)

ある周波数帯域において占有的に強いパワーを持つ周波数成分が存在する時、その帯域に含まれる周波数 bin において、IF はほぼ一定値となる。この値は、占有的な周波数成分のよい推定値を与える。音声において、このほぼ一定値となる IF は、基本周波数 F_0 の整数倍に位置する調波成分の周波数に相当する。図 1 に示すように、周波数 bin の中心周波数 ω_c と、各 bin の IF との関係をプロットすると、等間隔にならんだ階段状になる。この階段の水平部分と周波数 bin の中心周波数が一致する点 $\dot{\phi} = \omega_c$ が調波成分の周波数の推定値となり、この点のことを不動点とよぶ。隣り合った不動点の周波数の差が、F_0 の推定値を与える。

安部等は、振幅スペクトルの各成分 $S(\omega)$ を、対応する IF に基づき各周波数に再構成することで得られる IF 振幅スペクトル $g(\dot{\phi})$ を用いた F_0 推定法を提案している [9]。

$$g(\dot{\phi}) = \lim_{\triangle\dot{\phi}\to 0} \frac{1}{\triangle\dot{\phi}} \int_{\dot{\phi}<\dot{\phi}(\omega)<\dot{\phi}+\triangle\dot{\phi}} |\, S(\omega)\,|\, \mathrm{d}\omega.$$

通常の振幅スペクトルに比べて IF 振幅スペクトルでは、各調波成分がより鋭いピークを持つため、背景雑音下での F_0 推定に適しているとされている。しかし、IF 振幅スペクトルでは、電話音声のように基本波成分が低減する録音条件に対しては、推定性能が著しく低下するという問題がある。これは、振幅スペクトルが持つ周波数特性をそのまま利用している

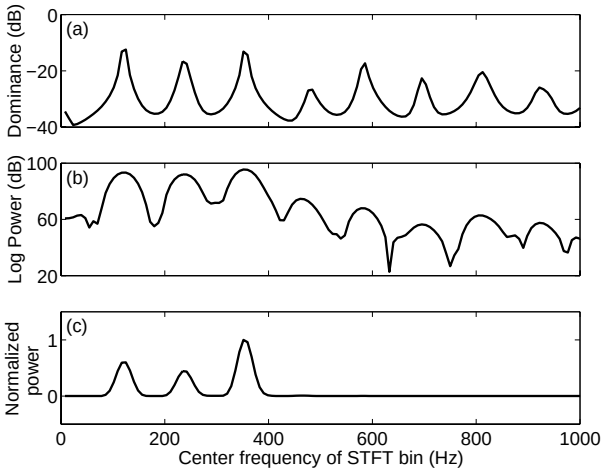

図 2　有声音 (背景雑音なし) の占有度スペクトル (a)、対数パワースペクトル (b)、およびパワースペクトル (c)

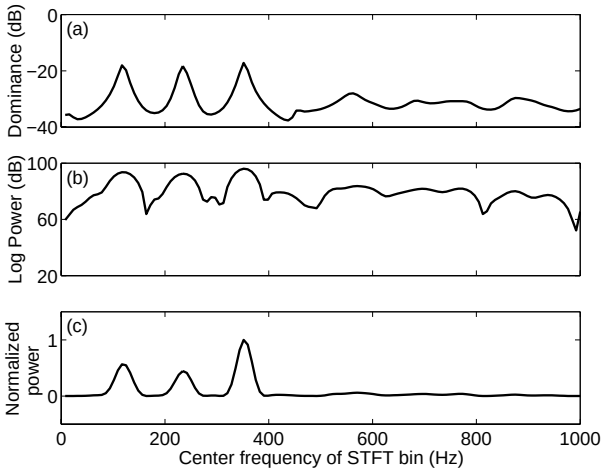

図 3　白色雑音下 (SNR:0 dB) での有声音の占有度スペクトル (a)、対数パワースペクトル (b)、およびパワースペクトル (c)

ためであり、振幅スペクトルにスペクトル変形が加わると、その影響を直接受けてしまうからである。また、上式の計算は、IF に関する微分処理を用いており、一般に、微分は雑音も強調する効果がある。このため、IF 振幅スペクトルに雑音抑制効果は期待できないことが予想される。

阿竹らは、Cohen の帯域幅方程式 [12] を用いて不動点の安定性を評価し、これに基づく F_0 推定法を提案している [10]。この方法は、低い SNR 条件でも高精度な F_0 推定を行うことができる。しかしながら、分析に用いる時間窓の長さを適応的に決めるために、事前に、F_0 の概算値を求める必要があった。しかも、F_0 推定の頑健さはこの概算値の精度に大きく依存するという問題があった。

本稿では、従来法の問題点を克服する新しい F_0 推定法を提案する。このため、入力音から背景雑音を抑制するとともに、周波数パワーの影響を正規化して取り除くことでスペクトル包絡が平坦な周波数特性を有する占有度スペクトルを定義する。以下、2 節で、占有度の定義および占有度を用いた F_0 推定法を構成し、3 節でその有効性を評価する。

2.　F_0 推定法

2.1　占有度の定義

我々は、各調波成分が近傍の周波数 bin の出力成分中をどの程度占めているかを評価するために、調波成分の占有度 (degree of dominance) $D_0(\omega_c)$ を以下で定義する。

$$D_0(\omega_c) = \log(1/B(\omega_c)^2), \tag{1}$$

$$B(\omega_c)^2 = \frac{\int_{\omega_c-\triangle\omega/2}^{\omega_c+\triangle\omega/2} (\dot{\phi}(\omega)-\omega_c)^2 S(\omega)^2 \mathrm{d}\omega}{\int_{\omega_c-\triangle\omega/2}^{\omega_c+\triangle\omega/2} S(\omega)^2 \mathrm{d}\omega}. \tag{2}$$

$B(\omega_c)^2$ は、中心周波数 ω_c を持つ周波数 bin の近傍 $(\omega_c - \triangle\omega/2 < \omega < \omega_c + \triangle\omega/2)$ の周波数 bin において、瞬時周波数 $\dot{\phi}(\omega)$ と ω_c の差を振幅スペクトル $S(\omega)$ の 2 乗で重み付き平均をとったものである。

図 1 に示したように、ある中心周波数 ω_c が、一つの占有的な周波数成分に対する不動点となる時、この ω_c と近傍の周波数 bin の $\dot{\phi}(\omega)$ はほぼ同一の値をとる。このため、式 (2) の $B(\omega_c)^2$ は極小値となる。式 (1) で、その逆数の対数をとり、同じ点で鋭いピーク値をとるようにしたものが占有度である。一方、ω_c が不動点であっても、対応する周波数成分が雑音の影響を強く受けている場合は、占有度は鋭いピークを持たない。これは、この不動点近傍の周波数 bin の IF は一定値とならず、各 bin の中心周波数とほぼ一致して増加し、ω_c と $\dot{\phi}(\omega)$ の差が大きくなるためである。この結果、ω_c が占有的な周波数成分の不動点に相当するときのみ、占有度は鋭いピークを持つことになる。なお、$B(\omega_c)^2$ の定義は、Cohen の帯域幅方程式の定義 $B^2 = \int(\omega-\bar{\omega})^2 S(\omega)^2 \mathrm{d}\omega / \int S(\omega)^2 \mathrm{d}\omega$ ($\bar{\omega}$ は平均周波数) と類似しているが、その目的と計算結果は大きく異なるものである (2. 5 節参照)。

2.2　占有度スペクトルとその有効性

占有度 $D_0(\omega_c)$ をすべての周波数 bin について計算し、周波数 bin の関数として表現したスペクトルのことを我々は占有度スペクトルと呼ぶ。以下で、占有度スペクトルを用いた F_0 推定法が、背景雑音のような加法的雑音と空間音響や電話伝送特性のような乗法歪みの両方に対して、耐久性が高い方法となることを示す。

2.2.1　加法的雑音に対する耐久性

図 2(a) に背景雑音なしの場合の有声音の占有度スペクトルの例を、対数パワースペクトル (同図 (b)) と

対比して示す。占有度スペクトルはすべての調波成分に対応して、対数パワースペクトルより鋭い周波数ピークを持っていることがわかる。これに対し、SN 比 0dB の白色雑音が加わった場合を図 3(a)(b) にそれぞれ示す。占有度スペクトル (図 3(a)) は、400Hz 以下の周波数では依然としてより鋭いピークを有しているのに対し、それ以上の周波数では明確なピークを持たず山と谷の比が小さい。これに対し、対数パワースペクトル (同 3(b)) も同様に、400Hz 以上でピークは無くなるが山と谷の比は占有度より大きい。このことは、占有度が調波成分に対応するピークは強調する一方で、背景雑音が大きい部分のスペクトル形状の山谷は平滑化する効果を持つことを示している。例えばケプストラム解析することを想定すると、このようなスペクトル特性は頑健な F_0 推定にとって有効であることがわかる。ところで、図 2(c) と 3(c) は、対数を取る前の線形のパワースペクトルを示している。両者のピーク部分の違いはほとんどなく、400Hz 以上で雑音成分による違いが若干あるだけである。すなわち、調波成分のピークと背景雑音のパワーの差は非常に大きい。これより、パワースペクトルでも背景雑音に対しては頑健な F_0 推定法が構成できることが予想される。

2.2.2 乗法歪みに対する耐久性

占有度スペクトルは、音声のフォルマントピーク等に由来するスペクトル包絡を白色化し、対数パワースペクトルやパワースペクトルと比べて平坦になるという性質を有する。占有度のこの性質は、スペクトル変形を正規化してその影響を低減する効果を持っている。例として、図 4 に同じ入力音声に SRAEN フィルタをかけた場合の各スペクトル表現を示す。SRAEN フィルタは ITU-T の勧告で推奨されている電話マイクの特性を模擬するフィルタである [11]。占有度スペクトル (図 4(a)) の第 1,2 ピークは図 2(a) と比較すると低くはなっているがその差はわずかである。これに対して、パワースペクトル (図 4(c)) の 300Hz 以下のピークは明らかに抑制されており、第 1 ピークについてはほとんど消失してしまっている。これは、パワースペクトルはスペクトル変形に対して極めて敏感であり、F_0 推定性能が劣化する原因となることを示している。なお、対数パワースペクトル (図 4(b)) も低域のピークは抑制されているが、そのピーク形状は明確に残っており、パワースペクトルほどはスペクトル変形の影響を受けない。

以上のことから、占有度スペクトルは、背景雑音とスペクトル変形の両方に対して頑健であり、一般の雑音環境下における F_0 推定に有用であるといえる。

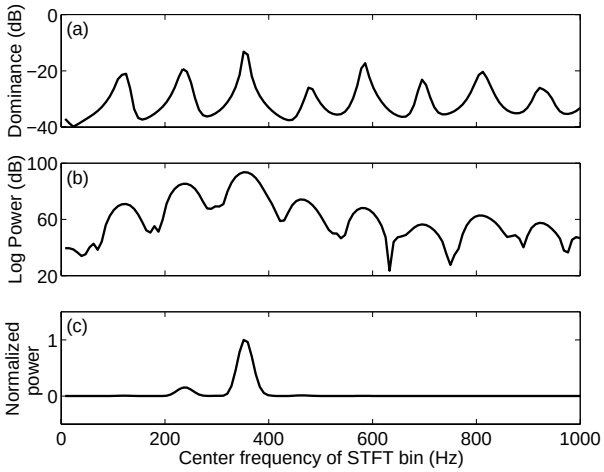

図 4 SRAEN フィルタをかけた有声音 (背景雑音なし) の占有度スペクトル (a)、対数パワースペクトル (b)、およびパワースペクトル (c)

2.3 調波構造占有度

占有度による F_0 抽出では、調波構造に関する占有度の総和 $D_{t0}(\omega_0)$ を調波構造占有度として定義する。式 (3) に示すこの値を最大化する ω_0 を F_0 候補周波数 (50〜500Hz) 中で求めることで F_0 を求める。

$$D_{t0}(\omega_0) = \sum_{l=1}^{n} \{D_0(\mathrm{r}(l\cdot\omega_0)) - \mathrm{E}(D_0(\omega_c))\}, \qquad (3)$$

$$F_0 = (1/2\pi)\ \arg\max_{\omega_0}\{D_{t0}(\omega_0)\}, \qquad (4)$$

ここで、l は ω_0 を基本周波数とみなしたときの高調波の次数であり、n は、最大 F_{max}(=1500 Hz) までに含まれる高調波の数である。また、$\mathrm{r}(\cdot)$ は $l\cdot\omega_0$ を最も近い周波数 bin の中心周波数 ω_c に変換する関数、$\mathrm{E}(D_0(\omega_c))$ は $D_0(\omega_c)$ の全周波数にわたる平均値である。

$\mathrm{E}(D_0(\omega_c))$ は、$D_{t0}(\omega_0)$ の期待値を 0 とすることで、半ピッチ・倍ピッチエラーを減少させるための項である。$D_{t0}(\omega_0)$ の期待値が 0 より小さいと、式 (3) は加算される高調波の数が少ないほど大きな値をとる傾向をもち、これと F_0 が高くなるにつれ F_{max} 以下に含まれる高調波の数が少なくなることから、結果的に、倍ピッチエラーが発生しやすくなる。また、期待値が 0 より大きい場合は、その逆の傾向があらわれる。

2.4 占有度による雑音に強い F_0 推定

2.4.1 F_0 推定フロー

図 5 に、占有度を用いた F_0 推定のフローを示す。まず、入力信号をダウンサンプリング (4 kHz) し、STFT (42ms ハニング窓、512 ポイント) により周波数領域信号に変換する。次に、各周波数 bin ごとに IF を求めるとともに、前節で説明した方法で占有度を計算する。IF の計算には、我々は、Flanagan の方法 [13] を

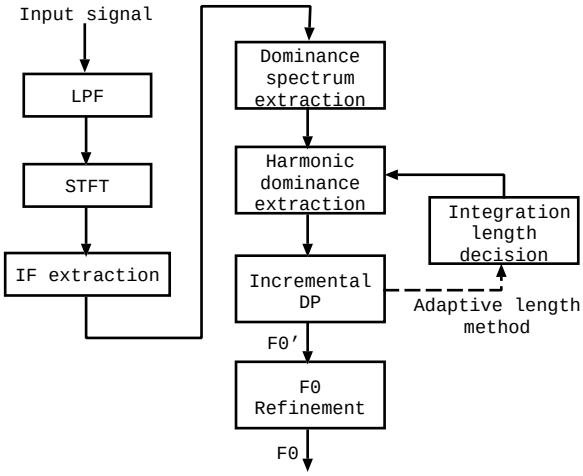

図 5　F_0 推定フロー

用いている。続いて、概算 F_0 を式 (4) を用いて求める。この時、F_0 が不連続に遷移する推定誤りを減少させるために、逐次 DP(後述) を用いている。最後に、2.5 節で説明する不動点の占有度に基づく方法で、更に精緻化した F_0 を求める。

2.4.2　逐　次　DP

我々は、連続的な調波構造占有度のピークを追跡して F_0 を求めるために、逐次 DP を用いる。従来、ケプストラム法などでは、連続的な F_0 を求めるために、通常の DP が用られてきた。しかし、我々は、F_0 抽出を様々な音声処理の前処理と想定しているため、DP のようなバッチ処理はあまり望ましくないと考える。このため、本稿では DP のアルゴリズムを改良して逐次処理を行うようにした逐次 DP を適用する。

逐次 DP では、各時刻において、すでに求められた現在時刻以前の調波構造占有度の時系列に対して、通常の DP を実行して現在の F_0 を求める。この方法で、過去から現在にわたる連続性について考慮した現在時刻の F_0 推定ができる。しかも、もともと DP は、実行途中において、現在時刻までの連続性を考慮した最適な F_0 パスを更新する逐次アルゴリズムであるため、逐次 DP にしても通常の DP と比べて余分な計算は発生しない。

2.4.3　積 分 範 囲

式 (2) を計算するために最適な積分範囲 ($\triangle\omega$ の値) は、安定な不動点近傍に見られる IF の水平部の周波数幅にほぼ一致し、このため、入力信号の F_0 に依存して決まる。したがって、F_0 の概算値を用いて最適な積分範囲を与えることができれば、より頑健に F_0 を推定することができる。我々は、より適切な積分範囲の決定法を与えるために、2 つの方法を導入する。1 つは事前情報に基づく固定積分範囲を用いる方法であり、もう 1 つは、事前情報を用いずに適応的に積分範囲を決定する方法である。

固定積分範囲を用いる方法では、事前情報として目的音声の話者の性別が事前にわかっているものとし、その性別の平均的な F_0 に対して最適な積分範囲を用いて、式 (2) を計算する。我々の実験では、男性の話者に対して最適な積分範囲は 130Hz 幅で、女性の話者に対しては 260Hz 幅であった。

一方、適応的な積分範囲を用いる方法では、事前情報を用いる代わりに、各入力フレームに対して、調波構造占有度を最大化する F_0 抽出を 2 回行う。つまり、1 回目に男性の話者にも女性の話者にも適用可能な固定の積分範囲を用いて F_0 の初期推定を行い、2 回目に F_0 初期推定値から求まる最適な積分範囲を用いてより精確な F_0 推定を行う。我々の実験では、初期推定のための積分範囲としては 260Hz 幅 (すなわち女性の話者に最適な積分範囲と同じ) を、また、詳細な F_0 推定には F_0 初期推定値の 67% ～110%の周波数幅を用いることで良好な推定結果が得られた。

2.5　占有度を用いた精確な F_0 推定

2.5.1　推定精度の改善

2.4 節の方法で求めた概算 F_0 近傍で、より精確な F_0 推定値を求めるために、我々は不動点の占有度に基づく F_0 精緻化法を導入する。大きな占有度を持つ不動点は、占有的な調波成分に相当する不動点であるため雑音の影響をあまり受けていないことが期待され、したがって、その IF は各調波成分の精確な周波数推定値を与えるものと期待される。この調波成分の周波数に基づき、それらを整数で割った値として F_0 を求めることでより精確な推定値を得る。

不動点の安定性を用いて F_0 を精緻化するという考えは、最初、阿竹らによって Cohen の帯域幅方程式を用いる方法として提案されている [10]。本稿では、占有度を用いることでこの方法を改良する。2.1 節で述べたように、占有度と Cohen の帯域幅方程式は関連性のある尺度であるが、不動点近傍の局所的な周波数領域における安定性を評価する尺度としては、占有度の方がより明確に定義された尺度である。(注1)

まず、F_0' を 2.4 節で議論した調波構造占有度の最大化によって求められる概算 F_0 とする。これに対し、精緻化した F_0 は以下のように定義される。

$$F_0 = \frac{1}{2\pi} \frac{\displaystyle\sum_{i=1}^{n} \sum_{\dot{\phi}\in\dot{\Phi}(i\cdot F_0')} (\dot{\phi}/i)\{D_0(r(\dot{\phi})) - c\}}{\displaystyle\sum_{i=1}^{n} \sum_{\dot{\phi}\in\dot{\Phi}(i\cdot F_0')} \{D_0(r(\dot{\phi})) - c\}}, \qquad (5)$$

(注1)：Cohen の定義は不動点まわりの IF の帯域幅を表現するものではなくて、パワースペクトルの周波数分布の中心 (=平均周波数) まわりのパワースペクトルの分布を表現したものである。さらに、Cohen の定義では、積分範囲を特定の周波数範囲に限定するための方法も提供されていない。

$$c = \min_{\dot{\phi}\in\dot{\Phi}(i\cdot F_0'),i=1\sim n} (D_0(r(\dot{\phi}))) + \epsilon. \quad (6)$$

ここで、$\dot{\Phi}(i\cdot F_0')$ は、概算 F_0 である F_0' の i 倍の周波数近傍 ($\pm 10\%$) に位置する不動点の IF の集合である。各 $\dot{\phi}(\in \dot{\Phi}(i\cdot F_0'))$ は、不動点から導かれる i 番目の高調波の周波数の候補であり、$\dot{\phi}/i$ は F_0 の候補を与える。この F_0 候補に関して、各不動点の占有度で重み付けした平均を取ったのが式 (6) の F_0 である。占有度の大きな値を持つ F_0 候補がより重み付けされることで、より精確な F_0 が得られるものと期待される。なお、c は、占有度を重みとして扱うためにすべての不動点に関する占有度が 0 以上になるようにするためのバイアスであり、ϵ は任意の小さな正の値でよい。

2.5.2 不動点の周波数補完

各不動点の IF は各周波数 bin の中心周波数間で周波数補完することで、より精確に求めることができる。ω_{c1} と ω_{c2} を隣り合った 2 つの周波数 bin の中心周波数とし、$\dot{\phi}_1$ と $\dot{\phi}_2$ をそれぞれの IF とする。これらの値が、次の式 (8) を満たす時、2 つの中心周波数の間に不動点が存在し、その IF $(=\dot{\phi})$ は、2 つの中心周波数を補完することで、以下のように定義することができる。

$$\dot{\phi} = \frac{(\omega_{c2} - \dot{\phi}_2)\omega_{c1} + (\dot{\phi}_1 - \omega_{c1})\omega_{c2}}{(\dot{\phi}_1 - \omega_{c1}) + (\omega_{c2} - \dot{\phi}_2)}, \quad (7)$$

$$\text{where } \dot{\phi}_1 > \omega_{c1} \text{ and } \dot{\phi}_2 < \omega_{c2}. \quad (8)$$

これにより、不動点の IF を連続値としてより精確に求めることができ、式 (5) で求まる F_0 も連続値となる。従来のケプストラム法などは求まる周波数が離散的であるため、求められた F_0 の時系列は階段状になっていた。これに対し、我々の占有度に基づく方法では特別な平滑化処理を行わなくても滑らかな曲線となる。

2.6 パワースペクトルを用いた F_0 推定

本節では、本稿で提案している占有度を用いた F_0 抽出法において、占有度の代用としてパワースペクトルを用いる方法を提案する。2.5 節で提案している F_0 の精緻化法は、パワースペクトルを用いる方法にも有効な方法である。2.1 節で述べたように、パワースペクトルでも、背景雑音と占有的な調波成分の差が対数パワースペクトルと比べて非常に大きくなるため、背景雑音下での高精度な F_0 推定を行える可能性がある。しかし、一方で、スペクトル変形に敏感であるため、録音条件によって性能が著しく変化することが予想される。

従来から、パワースペクトルを用いた F_0 推定法については、様々な応用の中で明示されないまま用いられてきたと考えられるが、その耐雑音性についてはほとんど議論されてこなかった。そこで、パワースペクトルを用いた F_0 推定法を構成し、次節の実験によりその性能比較を行う。

我々は、占有度による F_0 推定法のうち、式 (3), (5) および (6) の $D_0(\omega_c)$ を信号のパワーに置き換え、その他をそのまま用いることで、パワースペクトルによる F_0 推定法を構成する。ただし、信号のパワーはそのまま用いるよりも、そのスペクトル包絡成分を除去した残差スペクトル $R(\omega_c)$ を用いることで、次節の実験ではより良い結果が得られた。パワースペクトルから残差スペクトル $R(\omega_c)$ を求める計算は、対数パワースペクトルから高次ケプストラムに対応するスペクトルを求めるのと同様である。それは、パワースペクトル $S(\omega_c)^2$ に離散フーリエ変換を施し、その低域成分を 0 と置き換えた後に、離散逆フーリエ変換を施して、残差スペクトルを得た。

3. 実　　験

3.1 評価方法

目的音声の正解 F_0 を得る方法は、F_0 推定法の性能評価にとってとても重要である。従来、しばしば行われてきた方法では、雑音のない音声の対数パワースペクトルを人手で見て正解 F_0 を決定したり、計算機がいずれかの F_0 推定法を用いて推定した F_0 を正解 F_0 としていた。これらの方法の問題点は、人や採用した F_0 推定法がどのようにして正解 F_0 を決めるかによって、正解 F_0 にかたよりが生じ、そのかたよりによって評価すべき F_0 推定法の性能が左右されることである。また、どちらの方法も、観測された音声信号から正解 F_0 を求めているため、声道の周波数特性が正解 F_0 に影響を与えてしまっている可能性がある。我々にとっては、F_0 推定の目的の一つは、高性能 vocoder の音源推定であり、声道の影響をうけていない正解 F_0 が望ましい。

そこで、我々は、音声を録音するのと同時に採取した electro glottal graph (EGG) 信号から正解 F_0 を計算する [10]。EGG は声帯の開閉から直接取り出される信号であるので、正解 F_0 を計算するためにほぼ理想的な信号であるといえる。ただし、EGG 信号を用いても、推定法による正解 F_0 のかたよりの問題を完全に回避することはできない。これに対処するために、我々は、正解 F_0 として唯一の値を求めるかわりに、正解 F_0 と目的音声の F_0 を、評価の対象となる同じ F_0 推定法を用いて、EGG 信号と背景雑音下の音声からそれぞれ推定し、それらの違いを比較するという方法をとる。この方法は完全に誤りのない唯一の正解 F_0 を定義するものではないが、各 F_0 推

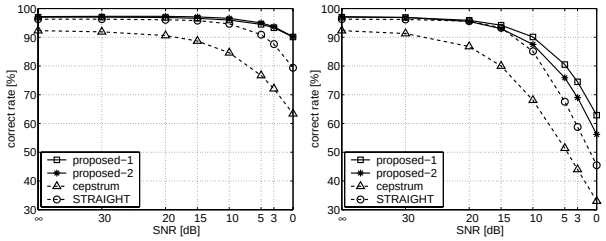

図 6　白色雑音下 (左図) およびマルチトーカ雑音下 (右図) での F_0 正解率

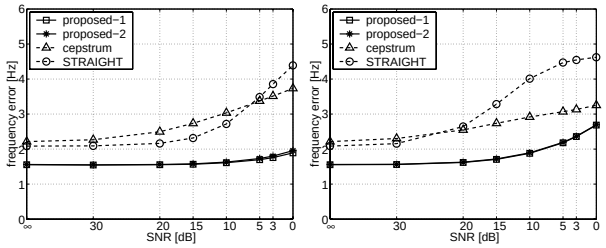

図 7　白色雑音下 (左図) およびマルチトーカ雑音下 (右図) での F_0 の実効誤差

定法ごとに、背景雑音および声道の影響に対する頑健性を評価するという観点で、正解 F_0 の推定法に依存しない評価を行うことができる。我々は、正解 F_0 と目的音声の F_0 の違いを、F_0 正解率と F_0 の実効誤差に基づいて評価する。F_0 正解率とは、目的音声の推定 F_0 が正解 F_0 の $\pm 5\%$ に入っている割合であり、F_0 の実効誤差とは、正解と判定された目的音声の F_0 推定値と正解 F_0 の RMS (root mean square) 誤差である。主に、前者は F_0 推定法の背景雑音に対する頑健性を、後者はその精度を評価する尺度である。

3.2　データベースおよび雑音

目的音声には男女各 2 名 (計 4 名) が発話した 30 種類の文 (計 120 文) からなる日本語発話データベース (16kHz 標本化、16bit 量子化) を用い、背景雑音には、白色雑音と、マルチトーカ雑音を用いた。マルチトーカ雑音は、上記発話データベースからランダムに選択した発話が、同時に 10 個混合するようにして作成したものである。各発話は、音声区間の平均パワーが一定になるようにパワーの正規化を行っている。例えば、マルチトーカ雑音下の目的音声を人間の耳で聞いた場合、0 dB SNR では目的音をほぼ聞き分けることができるが、−5 dB SNR では、約半分の音声区間が聞き分けられなくなり、−10 dB SNR では全く聞き分けられなくなる。

3.3　実験結果

3.3.1　占有度スペクトルによる方法

図 6 は、背景雑音下での F_0 正解率を示している。提案法で 2.4.3 節で述べた事前情報を用いる方法 (porpiosed-1)、事前情報を用いない方法 (proposed-2)、STRAIGHT で用いられている F_0 推定法 [1]、およびケプストラム法を比較した。図 6 は、∞ dB から 0 dB の各背景雑音に対して、proposed-1 と proposed-2 が従来法より正しく F_0 推定が行えていることを示している。特に、proposed-2 は事前情報を用いていないにもかかわらず、proposed-1 と同等の性能が得られている。proposed-1 は、10 dB SNR よりも大きなマルチトーカ雑音下の場合のみ、proposed-1 よりも優れた結果を出している。

図 7 は、同じ条件での F_0 推定における F_0 の実効誤差を示している。F_0 の実効誤差は、上記の F_0 正解率の測定時に、すべての方法が正解 F_0 を抽出できた区間に関して計算している。2 つの提案法の精度はほぼ同じであり、かつ従来法に比べて明らかに優れた結果を出している。

3.3.2　パワースペクトルによる方法

2.6 節で提案した、占有度の代用としてパワースペクトルを用いる方法を評価した。図 8 の太線は、この方法を用いた場合 (PowerSpec-1) の、背景雑音下での F_0 正解率を示す。結果は、85% くらいで一定で、図 6 の占有度を用いた方法と比較すると、特に、雑音が少ない (SNR が高い) 状況での性能が悪い。これは、EGG 信号から正解 F_0 を推定する際に誤差が大きく発生しているのが原因である。EGG 信号では低域の周波数パワーが、F_0 成分のパワーよりも大きくなることがあり、これが正解 F_0 の推定精度を劣化させていると考えられる。

そこで、EGG 信号の周波数特性の影響を取り除くために、事前に入力信号にプリエンファシスを適用する (PowerSpec-2)。この時の結果を、図 8 の細線で示す。雑音が少ない条件での正解率は改善しているが、逆に、雑音が強くなるにつれて急激に性能が悪化するようになってしまった。これは、プリエンファシスによって目的音声の F_0 推定の頑健性が損なわれてしまったのが原因である。

最後に、正解 F_0 は EGG 信号にプリエンファシスをかけて、目的音声の F_0 はプリエンファシスをかけずに推定した場合 (PowerSpec-3) の結果を図 8 の点線に示す。正解率は上記の 2 つより明らかに改善され、しかも、図 6 の proposed-1 をわずかにしのぐ結果が得られた。これらの結果は、パワースペクトルが背景雑音に対しては頑健であるが、スペクトル変形に対しては脆弱であることを示している。適切な前処理を選択することができれば、単純なパワースペクトルでも潜在的に頑健な性能を引き出せる可能性は持っているが、あらゆる録音条件に適応する前処理を特定することは困難である。これに対し、占有度は、前処理を適用しなくても、ほぼ最適に近い F_0 正解率を与えることができるといえる。

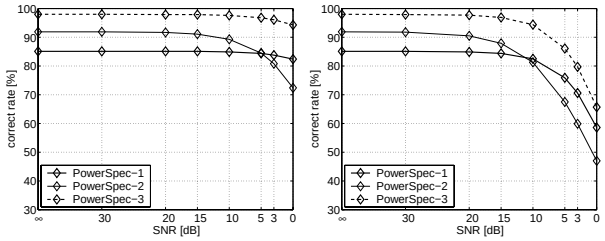

図 8　白色雑音下 (左図) およびマルチトーカ雑音下 (右図) での F_0 正解率

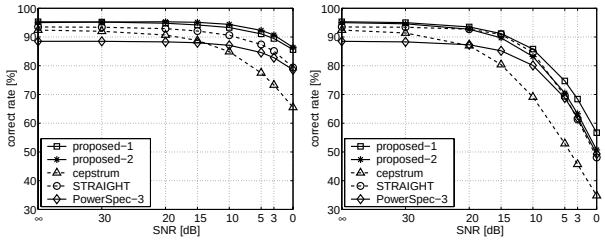

図 9　白色雑音下 (左図) およびマルチトーカ雑音下 (右図) で SRAEN フィルタをかけた場合の F_0 正解率

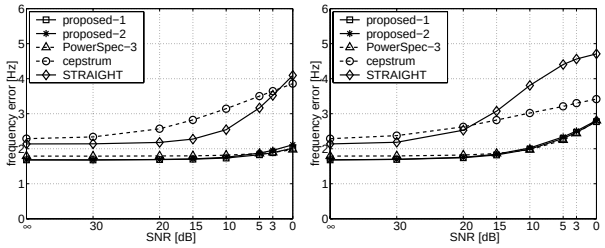

図 10　白色雑音下 (左図) およびマルチトーカ雑音下 (右図) で SRAEN フィルタをかけた場合の F_0 の実効誤差

3.3.3　スペクトル変形に対する頑健性

図 9、10 は、背景雑音下で目的音声に SRAEN フィルタをかけた場合の F_0 正解率と F_0 の実効誤差を示している。SRAEN フィルタは電話マイクの理想的な特性を模擬するもので、300Hz 以上の高域通過フィルタになっている [11]。図より、2 つの提案法は、SRAEN フィルタの影響下でも最善の性能を出していることがわかる。一方、PowerSpec-3 では、SRAEN フィルタに由来するスペクトル変形のために、F_0 正解率が著しく劣化していることがわかる。

4. ま と め

音声信号の調波成分が背景雑音から影響を受けていない度合いを評価するために調波成分の占有度を定義し、この占有度を用いて、背景雑音に加えてスペクトル変形が加わる条件下でも頑健で精度のよい F_0 推定法を提案した。実験により、提案法は、∞〜0 dB SNR の白色雑音下およびマルチトーカ雑音下で、F_0 正解率、F_0 の実効誤差の両方に関して、従来法より優れた性能を有することを示した。また、提案法は SRAEN フィルタに由来するスペクトル変形の影響に対しても最も頑健であることを示した。同時に、入力音声の周波数形状を注意深く補正することができれば、占有度の代用として単純なパワースペクトルを用いる場合でも、背景雑音に対して頑健な F_0 推定法が構成できることを示した。

今後の課題としては、カクテルパーティ状況下での音源分離などに応用するためには、F_0 推定法の頑健性の更なる改善が必要である。これには、例えば、複数音源の F_0 を同時に追跡するアルゴリズム [14] との統合や、マイクロフォンアレイなどを用いて得られる音源位置情報と組み合わせて用いる方法などが考えられる [3]。

謝辞 音声と EGG 信号のデータベースを提供いただいた和歌山大の河原教授に感謝する。片桐滋部長をはじめ、日頃より熱心な議論をいただいている NTT CS 研音声オープンラボの方々に感謝する。本研究の一部は CREST の支援を受けた。

文　献

[1] Kawahara, H. *et al.*, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communications*, Vol. 27, Nos. 3-4 pp. 187–207, Elsevier, 1999.

[2] Bregman, A. S., *Auditory scene analysis - the perceptual organization of sound*, MIT Press, 1990.

[3] Nakatani, T. *et al.*, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communications*, Vol. 27, Nos. 3-4, pp. 209–222, Elsevier, 1999.

[4] Hess, W., *Pitch determination of speech signals*, Springer-Verlag, Berlin, 1983.

[5] Rabiner, L. R., *et al.*, "A comparative performance study of several pitch detection algorithms," *IEEE trans. Audio Electroacoust.,* vol. ASSP-24, pp. 399–417, Oct., 1976.

[6] Cheveigne, A. D., *et al.*, "Comparative evaluation of F_0 extraction algorithms," *Proc. EUROSPEECH,* vol. 4, pp. 2451–2454, Sep., 2001.

[7] Liu, D., *et al.*, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE trans. Speech Audio Processing*, vol. 9, no. 6, Sep., 2001.

[8] Charpentier, F.J.,"Pitch detection using the short-term phase spectrum," ICASSP 86, Tokyo, 1086.

[9] 阿部 他, "瞬時周波数に基づく雑音環境下でのピッチ推定" 信学論 Vo.J79-D-II, No.11, pp.1771–1781, 1996.

[10] 阿竹 他, "調波成分の瞬時周波数を用いた基本周波数推定方法" 信学論 Vol. J83-D-II, NO.11, pp.2077–2086, 2000.

[11] "ITU-T Recommendation P.11", 1994.

[12] Cohen, L., *Time-frequency analysis,* Prentice Hall, 1995.

[13] Flanagan, J. L. *et al.*, "Phase vocoder," *The Bell System Technical J.,* vol. 45, pp. 1493–1509, 1966.

[14] 長渕 他, "混合音声における音声強調・抑圧" 信学論, Vol.62-A, No.10, 1979.