音楽情報科学 19－15
(1997. 2. 21)

IFスペクトログラム：音声信号の時間周波数表現の一手法

阿部 敏彦† 小林 隆夫† 今井 聖†

†東京工業大学 精密工学研究所

†〒226 横浜市緑区長津田町 4259

Email: tabe@pi.titech.ac.jp

あらまし 本論文では音声の時間-周波数表現の一手法として，IF(instantaneous frequency) スペクトログラムを提案する．IF スペクトログラムは従来の時間-周波数表現である短時間フーリエ変換（STFT）などに比べて，音声など準周期的信号の倍音構造を明確に表すものである．IF スペクトログラム上では倍音成分は非常に鋭いピークを持ち，倍音構造を明確に見ることができる．さらに，IF スペクトログラムに時間軸伸縮を導入する．その目的は，倍音周波数の急激な変化が引き起こす，スペクトログラム上での倍音成分どうしが重なり合ってしまうという現象が起きないようにすることである．フレームごとに倍音周波数の時間変化に従い時間軸伸縮を適用することにより，さらに明確な倍音構造を得ることが可能となる．

キーワード 時間-周波数表現，瞬時周波数，時間軸伸縮

THE IF SPECTROGRAM: AN APPROACH FOR TIME-FREQUENCY REPRESENTATION OF SPEECH

Toshihiko Abe†, Takao Kobayashi† and Satoshi Imai†

†Precision and Intelligence Laboratory, Tokyo Institute of Technology

†4259, Nagatsuta, Midori-ku, Yokohama, 226 Japan

Email: tabe@pi.titech.ac.jp

Abstract We propose a new spectral representation, the instantaneous frequency (IF) spectrogram, which more clearly shows the harmonic structure of quasi-periodic signals such as speech than conventional time-frequency (TF) representations such as the short time Fourier transform (STFT) spectrograms. In the IF spectrogram, harmonic components have very sharp spectral peaks and the harmonic structure is clearly seen. Next, we incorporate time-warping into the IF spectrogram to avoid overlapping of the harmonics in the spectrogram, that is caused by rapid change of the harmonic frequencies. The time-warping is applied to the signal frame by frame according to the change of the harmonic frequencies. As a result, we can obtain a spectrogram which shows sharp harmonic structures.

key words time-frequency representation, instantaneous frequency, time-warping

1. はじめに

音声信号の倍音構造の時間–周波数表現は，話し声や楽器音の分析に重要であり，この目的では通常，短時間フーリエ変換（STFT）やウィグナー分布（WD）[1]が使われている．それ以外にも時間変化を伴う信号の時間–周波数分析が多数提案されており，多くの応用において有効性が示されているが，多くは音声の倍音構造表現という目的においては多少の欠点を持つものと思われる．例えばSTFTでは倍音の位置のピークが窓関数のためにぼやけ，十分な時間–周波数解像度が得られないことがある．またWDではより良い解像度が得られるが，クロスタームを生じるという欠点を持つ．なお，WDにおけるクロスタームの問題を解決しようとする方法としては，RID[3]が提案されている．

本論文では倍音構造の分析のためのスペクトル表現として，音声のような準周期的信号の倍音構造を明瞭に表す，IF(instantaneous frequency) スペクトログラムを提案する．IF スペクトログラムは，IF 振幅スペクトル[4][5]の時間–周波数表現である．IF 振幅スペクトルは時間と周波数の関数として定義された瞬時周波数に基づいており，瞬時周波数を得るためにSTFTに基づくフィルタバンクを用いている．IF 振幅スペクトルはSTFT振幅スペクトルの積分の合計を保持するように瞬時周波数軸上に射影したものである．

我々はさらに以下に述べる問題を解決するために，IF スペクトログラムに時間軸伸縮の手法を導入する．IF スペクトログラムにおいても倍音成分の周波数が急速に変化するときは各成分を分離することが難しくなり，倍音成分どうしが周波数領域で重なり合いやすい．この現象を回避するために，倍音周波数の変化に合わせてフレームごとに時間軸伸縮を適用することを行なう．伸縮された時間軸上では，各分析フレーム内で倍音周波数は一定となるように観測され，そのため各倍音成分の分離がより簡単になる．倍音成分の分離のために時間軸伸縮を採用している他の手法としては文献 (9) があり，そこでは時間軸伸縮は基本周波数成分の位相に基づいて実現されている．一方，我々の手法は瞬時周波数とその時間導関数に基づいており，基本周波数つまりピッチの検出は行なっていない．

2. 時間–周波数平面上での瞬時周波数

信号 $x(t)$ の短時間フーリエ変換（STFT）は

$$X(\omega,t)=\int_{-\infty}^{\infty} x(\tau)w(\tau-t)\,\mathrm{e}^{-j\omega\tau}d\tau \tag{1}$$

により定義されている．ただし $w(t)$ は窓関数である．この $X(\omega,t)$ からフィルタバンク表現

$$F(\omega,t)=\mathrm{e}^{j\omega t}X(\omega,t) \tag{2}$$

を考えるとき，$x(t)$ は基底関数 $f(\omega,t)=w(t)\,\mathrm{e}^{j\omega t}$ の線形重ね合わせとして，

$$x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}F(\omega,\tau)f(\omega,\tau-t)d\omega d\tau \tag{3}$$

と表すことができる．ただし $\int_{-\infty}^{\infty}\{w(t)\}^2dt=1$ とする．従って $F(\omega,t)$ は基底関数 $f(\omega,t)$ の係数とみなすことができる[10]．

ここで点 (ω,t) における瞬時周波数を

$$\lambda(\omega,t)=\frac{\partial}{\partial t}\arg[F(\omega,t)] \tag{4}$$

と定義する．$F(\omega,t)=a+jb$ とおけば，瞬時周波数は

$$\lambda=\frac{a\frac{\partial b}{\partial t}-b\frac{\partial a}{\partial t}}{a^2+b^2} \tag{5}$$

により与えられる[11]．さらに

$$\begin{aligned}\frac{\partial}{\partial t}F(\omega,t)&=\frac{\partial a}{\partial t}+j\frac{\partial b}{\partial t}\\&=\int_{-\infty}^{\infty}\left(-\frac{\partial w(\tau-t)}{\partial t}+j\omega w(\tau-t)\right)\mathrm{e}^{-j\omega(\tau-t)}x(\tau)d\tau\end{aligned} \tag{6}$$

の関係が得られ，従って式（5）の $\partial a/\partial t$ と $\partial b/\partial t$ は式（1）の窓関数 $w(\tau-t)$ を $\left(-\frac{\partial}{\partial t}w(\tau-t)+j\omega w(\tau-t)\right)\mathrm{e}^{j\omega t}$ で置き換えることで求めることができる．

3. IF 振幅スペクトル

ここで簡単に IF 振幅スペクトル（瞬時周波数に関する振幅スペクトル）の定義[4][5]を示す．式（2）より $|F(\omega,t)|=|X(\omega,t)|$ となるので，$|F(\omega,t)|$ はSTFT振幅スペクトルに等しくなる．ここである固定された時間 t において，周波数軸 ω 上の微小区間 $\lambda_0\le\lambda(\omega,t)\le\lambda_0+\Delta\lambda$ で $|F(\omega,t)|$ の積分をとり，その積分値を微小区間に対応する振幅とみなす．そして次式のように $\Delta\lambda\to 0$ の極限をとり，その振幅の極限値を IF 振幅スペクトルと定義する．

$$g(\lambda_0,t)=\lim_{\Delta\lambda\to 0}\int_{\lambda_0\le\lambda(\omega,t)\le\lambda_0+\Delta\lambda}|F(\omega,t)|d\omega. \tag{7}$$

式（7）は瞬時周波数に関する時間–周波数表現とみなすこともできるので，$g(\lambda,t)$ を IF スペクトログラムと呼ぶ．図1に STFT スペクトログラム $|X(\omega,t)|$ と IF スペクトログラム $g(\lambda,t)$ の比較を示す．音声データは標本化周波数 10kHz，12 ビットで量子化したものを用いた．また式（1）を離散時間信号に対して計算するために，512点FFTを 2msec 間隔で実行した．窓関数 $w(t)$ には，長さ 40msec のブラックマン窓を用いた．これらの条件はSTFT スペクトログラムと IF 振幅スペクトログラムにおいて同じである．

図1(a) の STFT スペクトログラムと比較して図1(b) の IF スペクトログラムでは倍音周波数が非常に明瞭な線として表れていることが分かる．IF スペクトログラムのこのような特性は，定義した瞬時周波数が倍音周波数上の点に集中しやすいという特性によるものである．

4. 時間軸伸縮

倍音の分析では，各倍音成分が互いに分離され，各成分を抽出できるようになることが望ましい．しかしピッチが急激に変化するときは，急激な周波数変化により帯域が広がるために，各倍音成分が互いに重なり合いやすくなり，そのため時間–周波数平面上で各成分を分離するのが難しくなる．また高い次数の倍音成分ほど，重なり合いは著しくなる．この現象は，例えば図1で倍音周波数が急激に変化する部分，つまり図の左上と右上付近に見られる．この問題を解決するために，以下のように周波数変化に従って信号の時間軸伸縮を適用する．

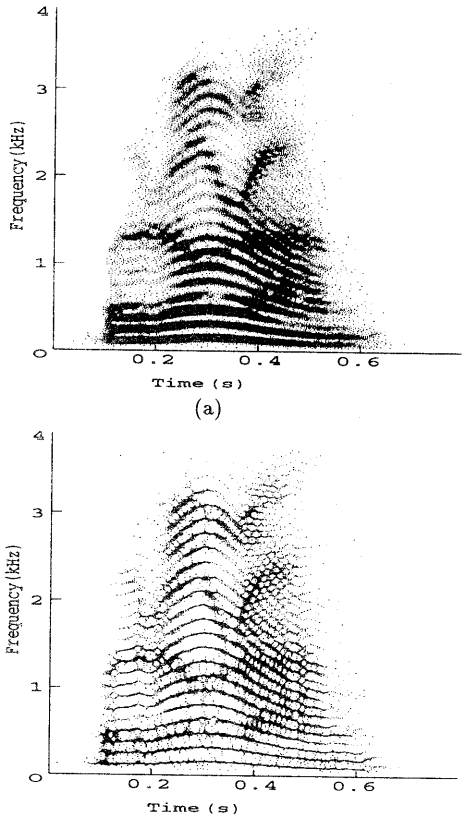


図 1: 男性話者「室蘭」のスペクトログラム（時間軸伸縮なし）(a)STFT スペクトログラム (b)IF スペクトログラム

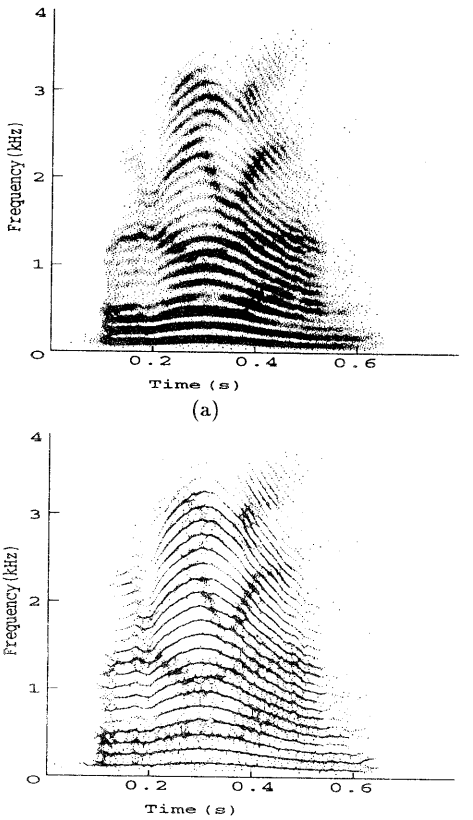


図 2: 男性話者「室蘭」のスペクトログラム（時間軸伸縮を導入）(a)STFT スペクトログラム (b)IF スペクトログラム

4.1 原理

まず，時間軸伸縮された後の時間軸を，時間軸 u と名付ける．また p を 2 つの時間軸の関係

$$u = p(t) \quad \text{and} \quad t = p^{-1}(u) \tag{8}$$

を決める伸縮関数とする．ただし $p(t)$ は連続で単調増加であると仮定する．時間軸 u は，その時間軸上での倍音周波数の変化が消えるように設定する．ここで各倍音成分を，振幅変調と周波数変調を受けた正弦波として記述する．つまり

$$x(t) = r(t) \cdot \cos(\theta(t)) \tag{9}$$

ただし

$$\theta(t) = \int^t \lambda_s(\tau) d\tau + \theta_0 \tag{10}$$

また，$\lambda_s(\tau)$ と θ_0 はそれぞれ瞬時周波数と初期位相である．そして u の関数であるということを表すのに添字 u を用いることにする．すると時間軸 t と u での位相の表現は次式のようになる．

$$\theta_u(u) = \theta_u(p(t)) = \theta(t). \tag{11}$$

すると u 軸上での瞬時周波数が一定となるための条件は

$$\left(\int^t \lambda_s(\tau) d\tau + \phi_0 \right) = \psi u + \psi_0 = \theta_u(u) \tag{12}$$

となる．ただし定数 ψ と ψ_0 はそれぞれ周波数と初期位相である．式（12）を u で 2 回微分し$\lambda_s(t) = \partial\theta/\partial t$ を代入すると次式を得る[6]．

$$\frac{d^2t}{du^2}\frac{\partial\theta}{\partial t} + \left(\frac{dt}{du}\right)^2 \frac{\partial^2\theta}{\partial t^2} = 0. \tag{13}$$

従って位相 $\theta(t)$ が与えられたとき，時間軸伸縮関数 $u = p(t)$ は微分方程式（13）の解の一つとして与えられる．

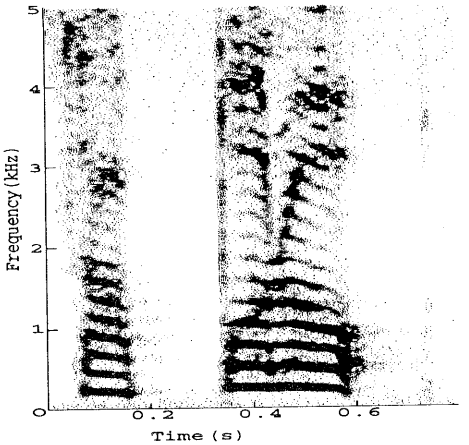
式（13）は次のように書き換えられる．

$$\frac{d^2u}{dt^2} = q(t)\frac{du}{dt} \tag{14}$$

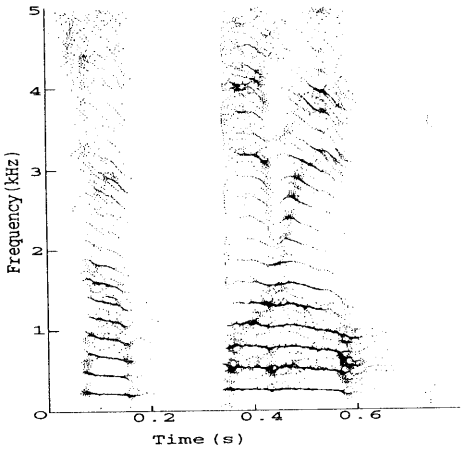
ただし

$$q(t) = \lambda_s(t)^{-1}\frac{\partial\lambda_s(t)}{\partial t} \quad \text{and} \quad \lambda_s(t) = \frac{\partial\theta(t)}{\partial t}. \tag{15}$$

ここで式（15）は $\lambda_s(t)$ の関数なので，t と u が満たす微分方程式（14）は$\lambda_s(t)$ つまり t 軸上の瞬時周波数により与えられることが分かる．

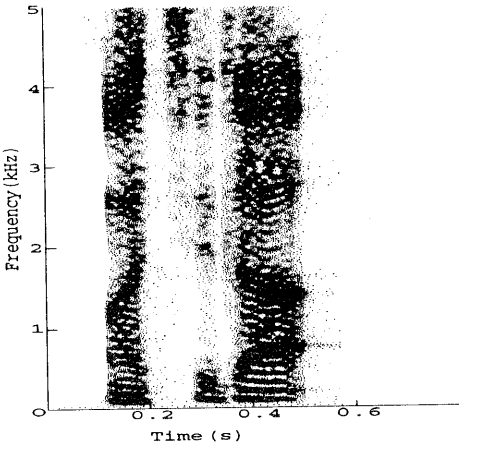


(a)

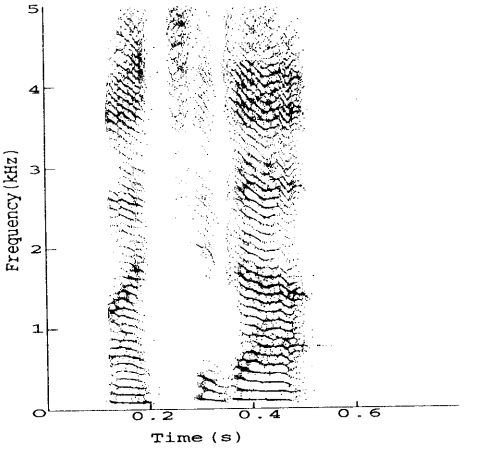


(b)

図 3: 女性話者「札幌」(a)STFT スペクトログラム (b)IF スペクトログラム



(a)



(b)

図 4: 男性話者「町田」(a)STFT スペクトログラム (b)IF スペクトログラム

4.2 時間軸伸縮の実現

実際の分析では，注目する時間 t_0 を瞬時周波数分析に用いる窓関数の中心点として，その近傍において時間軸伸縮関数 $u = p(t)$ を設定する，つまり時間軸伸縮を局所的に行なうことにする．計算を簡単にするため信号はチャープ，つまり周波数は時間に対して線形に変化するものと仮定する．

チャープ信号の周波数は時間の 1 次多項式となる．その結果として時間軸伸縮関数 $u = p(t)$ は t の 2 次多項式となる．伸縮された時間軸 u は時間 t_0 が変わるたびに設定し直す．つまり異なる値の t_0 について異なる u が対応する．そこで u 軸が t_0 におけるものであることを $u_{t_0} = p_{t_0}(t)$ と表す．他の変数についても同様にする．ここで $p_{t_0}(t_0) = 0$ とおけば，$p_{t_0}(t)$ は t_0 の近傍で定数 a と b により

$$p_{t_0}(t) = \frac{1}{2}a_{t_0}(t - t_0)^2 + b_{t_0}(t - t_0) \tag{16}$$

と書ける．そして $t = t_0$ となる点が u_{t_0} 軸の原点に対応し，その点上では両者の時間軸の尺度が同一，つまり $du_{t_0}/dt = 1$ であると仮定する．前者の仮定を式で表せば $u_{t_0} = 0 \leftrightarrow t = t_0$ となり，また後者は

$$b_{t_0} = \left.\frac{du_{t_0}}{dt}\right|_{u_{t_0}=0} = 1, \quad a_{t_0} = \left.\frac{d^2u_{t_0}}{dt^2}\right|_{u_{t_0}=0} = q(t_0) \tag{17}$$

となる．

ところで，時間軸伸縮を適用したときに式 (3) は

$$x(t) = \frac{1}{2\pi}\int\int F(\omega,\tau)f_{t_0}(\omega,\tau - t)d\omega d\tau \tag{18}$$

に変わる．ただし $f(\omega,t)$ は t_0 の関数となる．つまり

$$f_{t_0}(\omega,t) = (1 + a_{t_0}t)f(\omega,\frac{a_{t_0}}{2}t^2 + t) \tag{19}$$

となる．従って伸縮された時間軸上ではフーリエ基底は正弦波から作られているが，それは元の時間軸からは，チャープ信号で作られているように見える．

4.3 分析例

図2はSTFTスペクトログラム $|X(\omega,t)|$ とIFスペクトログラム $g(\lambda,t)$ の例であり，両者とも時間軸伸縮が適用されている．ところで時間軸 u を決定するためには式（15）の $q(t)$ の推定値を得る必要があるが，その方法は後に5章で述べるIFアトラクタに基づいている．時間軸伸縮は信号を標本点間において補間を行ない，時間軸伸縮関数に従って再度標本化することにより実現する．図1の時間軸伸縮なしの場合を見ると，周波数が急激に変化する所では倍音どうしが完全には分離されず，互いに重なり合っていることが分かる．一方図2の時間軸伸縮が適用されている場合には，倍音はほぼ完全に分離されている．図3と図4は，別の信号に対して時間軸伸縮を用いたIFスペクトログラムを求めた例である．

5. 時間–周波数平面上でのIFアトラクタ群

5.1 定義

IFスペクトログラム以外にも，時間–周波数上の瞬時周波数に関する幾何学的な性質として重要と思われるものがあり，それをIFアトラクタと呼ぶことにする．IFアトラクタとは簡単に言えば，倍音周波数の軌跡を示すものである．

関数 $\mu(\omega,t)$ を次式で定義する．

$$\mu(\omega,t) = \lambda(\omega,t) - \omega. \tag{20}$$

ここで (ω,t) 平面上で，次の条件が成り立つ点の集合を考える．

$$\mu(\omega,t) = 0 \quad \text{and} \quad \frac{\partial \mu}{\partial \omega} < 0 \tag{21}$$

これらの点は時間–周波数平面上で，倍音周波数の位置に対応する曲線の集合を構成する．またIF振幅スペクトルは一般に，これらの曲線群上に集中する傾向があるが，このことはIF振幅スペクトルが倍音周波数上で強いピークをもつという事実に一致するものである．従ってこれらの曲線群はあたかも瞬時周波数を引きつけるように振舞うので，「IFアトラクタ」と呼ぶことにする．また，それらは著者らが提案した瞬時周波数に基づく，倍音追跡フィルタバンクシステム[7][8]においてフィルタの中心周波数の軌道を引きつける性質を持つことも，それらを「アトラクタ」と呼ぶ理由の一つである．

5.2 倍音周波数変化量の抽出

前述した時間軸伸縮を実行するためには，倍音の時間変化に関するパラメータとして，式（15）の $q(t)$ を推定する必要がある．IFアトラクタは近似的に倍音周波数とみなすことができるので，$q(t)$ をIFアトラクタから推定することにする．これは次のように，アトラクタの時間導関数を得ることで実現される．

アトラクタの1つを時間の関数として，

$$\omega = \lambda_a(t) \tag{22}$$

のように記述する．式（21）と式（22）を合わせ，さらに t で微分すると，

$$\frac{d\lambda_a}{dt}\frac{\partial \mu}{\partial \omega} + \frac{\partial \mu}{\partial t} = 0 \tag{23}$$

を得る．これは次式のようにも書ける．

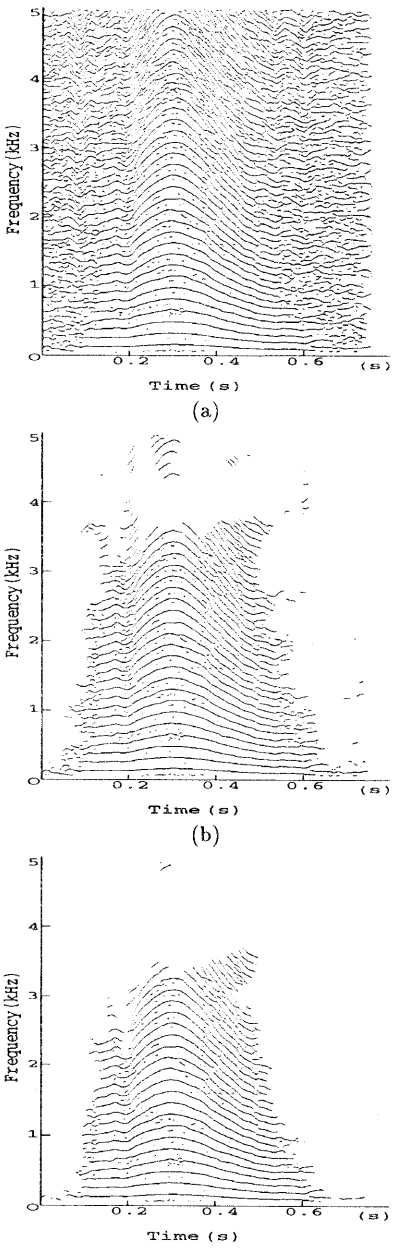


図5: 図1、図2と同じ信号での (a) すべてのアトラクタ (b) アトラクタのうち瞬時振幅がある閾値を越えているもの (c)(b) よりも閾値が大きい場合

$$\text{grad}\ \mu \cdot [\frac{d\lambda_a}{dt}, 1]^{\mathrm{T}} = 0. \tag{24}$$

ここで $[d\lambda_a/dt, 1]^{\mathrm{T}}$ は $\lambda_a(t)$ の接線ベクトルであることに注意すれば，(22) が示す曲線は μ の勾配に直角に交わるということを意味していることが分かる．このことは式（21）を満たす (ω,t) の集合は曲線上で μ が一定とな

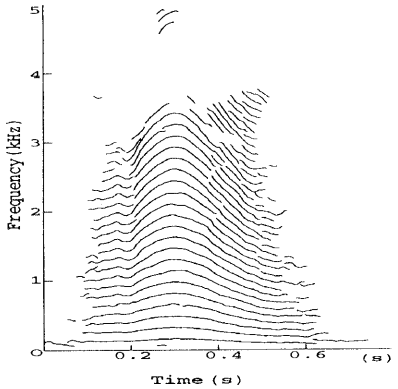


図 6: 推定された倍音

るような曲線を構成するということに対応する．

式（23）より，アトラクタの時間導関数は

$$\frac{d\lambda_a}{dt} = -\frac{\partial\mu}{\partial t}\Big/\frac{\partial\mu}{\partial\omega} \tag{25}$$

となる．これを倍音周波数の推定値として用いる．ここで次のように置く．

$$q_F(\omega,t) = \lambda(\omega,t)^{-1}\frac{d\lambda_a(\omega,t)}{dt} \tag{26}$$

ただし，

$$\frac{d\lambda_a(\omega,t)}{dt} = -\frac{\partial\mu}{\partial t}\Big/\frac{\partial\mu}{\partial\omega} \tag{27}$$

として，式（25）を (ω,t) のすべての点に対して定義する．

次にアトラクタ群に関して，次式のように周波数区間 $[\omega_0,\omega_1]$ で式（26）の平均を取る．

$$q(t) = \frac{\int_{\omega_0}^{\omega_1}\gamma(\omega,t)\alpha(\omega,t)|F(\omega,t)|q_F(\omega,t)d\omega}{\int_{\omega_0}^{\omega_1}\gamma(\omega,t)\alpha(\omega,t)|F(\omega,t)|d\omega} \tag{28}$$

ただし (ω,t) が式（21）を満たすとき $\gamma(\omega,t)=\delta(\omega)$ であり，それ以外では $\gamma(\omega,t)=0$ とする．ここで $\delta(\omega)$ はディラックのデルタ関数である．関数 $\gamma(\omega,t)$ はパラメータの抽出がアトラクタ上の点についてのみ行なわれるように機能する．式（28）の $q(t)$ を式（15）の推定値として用いる．式（28）において，重み付け $\alpha(\omega,t)|F(\omega,t)|$ はIF振幅スペクトルの正規化された局所的2次モーメントの関数であり，(ω,t) における成分の周期性の強さを示すものである[4][5]．$\alpha(\omega,t)$ を重み付けとして用いる理由は，アトラクタ群全体は倍音周波数だけでなく雑音的成分のの周波数にも対応しており，また倍音成分は雑音成分より周期性が強いはずなので，周期性を表す $\alpha(\omega,t)$ に比例した強調をアトラクタに施すためである．

5.3 分析例

図5にIFアトラクタの例を示す．まずアトラクタ群を時間軸伸縮なしで求め，式（28）により $q(t)$ の推定値を得る．ここで式（28）の周波数区間 $[\omega_0,\omega_1]$ は，時間軸伸縮なしでもアトラクタを用いて倍音周波数が大体正しく推定される範囲として，$\omega_0 \simeq 160$Hz，$\omega_1 \simeq 1$kHz のように設定する．そして推定された $q(t)$ を 40ms のハニング窓によるローパスフィルタに通し雑音的な変動を取り除く．そして時間軸伸縮を適用し，再度分析を行なうことになる．図5は時間軸伸縮を適用したときのアトラクタを示している．図5(a)は音声のすべてのアトラクタ群を表している．図5(b)と(c)はアトラクタ群のうち，瞬時振幅が前もって異なる値に設定した閾値を越えているものを表している．図6は倍音を推定したものである．ここでは倍音推定の一手法として，アトラクタ群のうち瞬時振幅が定められた振幅の閾値を越えている部分で，その時間長が定められた時間長の閾値よりも長いものとして求めている．

6. おわりに

我々は音声の時間-周波数表現の一手法としてのIFスペクトログラム，また倍音周波数の軌跡を表すIFアトラクタを提案した．さらにこれらの分析能力を向上させる手段として，時間軸伸縮の手法を示した．そしていくつかの分析例で，IFスペクトログラムがSTFTスペクトログラムよりも音声の倍音構造を明確に表すことを示した．本論文の手法が人間の音声だけでなく，楽器音や混合音声や雑音付加音声の分析にも有効であるものと期待している．今後の課題は，IFアトラクタに基づく周波数推定を用いた音声合成と音声変換などである．

文 献

(1) M.D. Riley, *Speech time-frequency representations,* Kluwer Academic Publishers, 1989.

(2) M. Cooke and M. Crawford, "Tracking spectral dominances in an auditory model," in *Visual Representations of Speech Signals*,pp.197–204, John Wiley & Sons Ltd, 1993.

(3) W.J. Williams, "Reduced interference distributions: Biological applications and interpretations," *Proc. of IEEE,* vol. 84, No.9, pp. 1264–1280, Sep., 1996.

(4) 阿部敏彦，小林隆夫，今井聖，"瞬時周波数に基づく雑音環境下でのピッチ推定" 信学論，vol. J79-D-II, No.11, pp.1771–1781, Nov., 1996.

(5) T. Abe, T. Kobayashi and S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," *in Proc. ICSLP 96 (Philadelphia)* pp. 1277-1280, 1996.

(6) 阿部敏彦，小林隆夫，今井聖，"時間軸伸縮を導入した瞬時周波数に基づく倍音推定" 信学技法，SP94-94, pp. 23–28, 1995.

(7) T. Abe, T. Kobayashi and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency, " *in Proc. ICASSP 95 (Detroit)* pp. 756-759, 1995.

(8) T. Abe, T. Kobayashi and S. Imai, "Harmonics estimation based on instantaneous frequency and its application to pitch determination," *IEICE Trans. Information & Systems,* vol. E78–D, No.9, pp. 1188–1194, 1995.

(9) R. Kumaresan and C.S Ramalingam, "On separating voiced-speech into its components," *in Proc. Twenty-Seventh Asilomar Conference on Signals, Systems & Computers*, Pacific Grove. CA, vol. 1–2.

(10) P.P. Vaidyanathan, *Multirate Systems and Filter Banks*, Prentice-Hall, New Jersey, 1993.

(11) J.L. Flanagan and R.M. Golden, "Phase vocoder," *Bell Syst. Tech.*, vol. 45, pp. 1493–1509, 1966.