自動採譜システムの構築

Transcription System

電気通信大学大学院情報システム学研究科

情報ネットワーク学専攻

学籍番号：9951018

杉浦 勇樹

指導教官

阪口 豊

出澤 正徳

磯野 春雄

2001年2月1日 提出

目次

第1章　はじめに

1.1　自動採譜とは

自動採譜システムとは、CD等の実音響信号から

1. 何の楽器が
2. いつ
3. どの音の高さを
4. どれくらいの強さで

演奏されたかをシンボルとして出力するシステムである。

このようなシステムが完成すれば、実音響信号のデータベース化に非常に役に立つと考えられる。例えば「サビのフレーズのみ覚えているが、曲名を忘れた」という場合を考える。検索をかけるためにはその曲の音の高さ、タイミング等をシンボルとして保存しておく必要があるが、これらの情報をすべてデータベース化の為に手入力等を行うのは既に存在する楽曲の数を考えると不可能である。

1.2　本研究の対象

ジャズ音楽では、即興演奏というのが重視されているので、毎回演奏内容が異なる。これは、演奏の数だけ譜面が作れることを意味する。市販されている譜面もあるが一部の非常に有名な演奏のみである。よって音響信号から譜面の情報を生成したいという要求はより強いと考えられる。そこで本研究ではジャンルをジャズ音楽を対象とする。

また、ジャズ音楽の演奏技術を磨く上で既存の演奏を模倣するという練習法が重視されている。ジャズ音楽の譜面はあまり市販されておらず、プレイヤーは自分で音響信号から譜面を作成しており、このための労力は大きい。このことからも自動採譜システムへの要望は強いと考えられる。

ジャズ音楽一般を扱おうとすると、楽器数や同時発音数が多いため非常に難しくなる。ピアノトリオは楽器数も比較的少なく、楽器の種類も限定される。楽器数は少ないながらもリズム、ベース、コード、メロディーという音楽を構成する要素を備えているので、この編成における問題解決のノウハウは問題を広げた場合でも活かせると考えられる。第一段階として本研究ではピアノトリオを対象とする。

1.3　過去の研究

自動採譜システム全体の研究としては既に多くの研究がなされているが、それらが対象にしている音源の性質によって難易度が異なる。

単音の音響信号を対象とする自動採譜システムはほぼ100%の精度を実現しており、既に

市販されている。

複音、単一楽器を対象する自動採譜システムについては、青野らによりアコースティックピアノをもちいたセッションシステム [6] の中で使われている。周波数解析部には、Mult-BandFFT を用いており、Sub Harmonic Summation を用いて単音、もっとも有力なルートと5度のペアを探すことによってルート音を決定する手法により和音の同定を行っている。和音の精度はピッチ周波数2700cent～500cent の間で100%と唱っているものの、評価の対象に使っている和音の構成音として何を用いているのか、実際に当てている情報は和音の構成音なのか和音名なのかは不明である。

複音、複数楽器を対象とする自動採譜システムについては、柏野によって立ち上がり時刻、高調波成分の周波数のずれ等を用いた音色判断、ベイジアンネットワークを用いて周波数成分レベル、単音レベル、和音レベルの情報を統合している。各時刻間に和音遷移確率を事前知識として用い、精度を挙げている [12]。評価は3声の室内楽の演奏を使って行われており、一番高いものでも78%程度である。

また後藤 [10] [11] は市販の CD によるジャズやポピュラー音楽の音響信号から、リアルタイムでメロディーとベース音の音高推定を行うシステムを構築している。FFT フィルタバンクを用いて周波数解析と瞬時周波数を用いた補正を行った後、高調波構造を持つ音をモデル化した確率分布の重み付き和によってスペクトルのパワー分布が作られたものとし、それを EM アルゴリズムを用いた最尤推定により基本周波数の確率分布を求め、その分布をペナルティと信頼度を持つ複数のエージェントにより追跡し、出力を決定している。メロディーについては76%～95%, ベース音については65.8%～91.2%の検出率である。

自動採譜の観点から見たジャズのピアノトリオは

1. 複数楽器を含むアンサンブルで同時発音数が不定である
2. 打楽器が入っている
3. 中音域で音価の短い和音が弾かれる

という特徴がある。

このように複数楽器で打楽器が細かいリズムを刻んでいるような楽曲を対象にする研究は、上記の後藤による研究 [10] [11] 以外は特に行われておらず、しかも上記の研究でもメロディーとベースラインのみを当てるというもので、本研究のような伴奏パートまであてるような研究ではない。

1.4 本論文の構成

2章では、楽音の性質を述べ、そこから考察を進めピッチ周波数を求める方法を提案し、その評価を行う。

3章では、2つのノートが近接して鳴っているときにそれらをうなりを用いて識別する手法について提案を行いその有効性を探る。4章でまとめとする。

第2章　音の高さを求める

本章では、まず楽音の性質を述べる。そこから考察をすすめ、音の高さを検出する手法 Multeapics を新たに提案する。

2.1　楽音の性質

楽音は、基本周波数に対してその整数倍の高調波成分を含む。また、2つの純音が合わさった場合、それらは単純に和が取られるのではなく、それらの位相に応じて計算されるので、パワーとしては逆に弱まることもある。

基本周波数が欠けている、または非常に小さいパワーでもピッチがその基本周波数にあたる高さで知覚される、という現象がおこる。

また、楽器によっては基本周波数に対して高調波の周波数が多少ずれる、という性質もある。例えば、ピアノ音の場合 第15部分音の周波数が基本周波数の16倍に相当する [9]。

2.2　ピッチ周波数推定法

前章で述べたような構造をもつ楽音のスペクトルからピッチ周波数を見つけ出すために、各周波数毎に定義されるピッチらしさ (PitchLikelhood:$PL(x)$ と表す) を「$PL(x)$ は分布であり、そのピークの最大値がピッチ周波数を表す」と定義し、そのピークの最大値を求める問題と考える。

$PL(x)$ はスペクトル上のピーク周波数の最大公約数を求めれば良さそうである。しかし、

1. 対象が単音の楽音ではなく複数の値を取る可能性があること
2. 高調波の周波数にはずれがあること

という理由から、完全に最大公約数を表す分布とは定義できず、ある程度の誤差を吸収できる必要がある。

この処理を行っている手法に、Subharmonic Summation がある。

2.2.1　Subharmonic Summation

[7] では「ハーモニックサメーションとは高調波に散らばったエネルギーを基本周波数に集める処理に相当する」としている。また、文献 [3] では、Subharmonic Summation という言葉は用いていないものの、これと同等の処理を行っている。さらに文献 [6] では、和音の構成音の同定に利用している。本研究では、文献 [3] で用いている手法を Subharmonic Summation と呼ぶ。

周波数 x cent に対し、パワースペクトルを $F(x)$ とおくと、Subharmonic Summation

は、ピッチ周波数らしさ $PL(x)$ を

$$PL(x) = \sum_{n=1}^{N_h} F(x + \log n) \tag{2.1}$$

としている。但し cent は周波数の単位で、周波数 f Hz との対応は $f = 440 \times \log_2 \frac{x-4800}{1200}$ である。

Subharmonic Summation が最大公約数らしさを表現できる理由を考える。スペクトル分布がその時刻におけるピークの振幅とその漏洩から表現できると仮定すれば、N_p 個のピークが存在し、i 番目のピーク周波数を c_i、パワーを p_i と表し、式 (2.1) は

$$PL(x) = \sum_{n=1}^{N_h} \sum_{i}^{N_p} p_i |\tilde{w}(c_i - (x + 1200 \log_2 n))|$$

と書ける。ここで、$\tilde{w}(x)$ は 周波数解析時に用いた窓関数 $w(x)$ に対応するスペクトルの漏洩を表す。この式は

$$\begin{aligned} PL(x) &= \sum_{i}^{N_p} p_i \sum_{n=1}^{N_h} |\tilde{w}(c_i - (x + 1200 \log_2 n))| \\ &= \sum_{i}^{N_p} p_i \sum_{n=1}^{N_h} h((c_i - 1200 \log_2 n) - x) \end{aligned}$$

と変形できる。ただし、$h(x) = \tilde{w}(x)$ とおいた。この式は、周波数 c_i に対して $\log n$ で低い周波数におけるピッチ周波数らしさを強めていると解釈できる。c_i は対数スケールなので、リニアスケールで考えれば、n 分の一の周波数のピッチ周波数らしさを強めていることになる。

最大公約数は公約数の特殊な値であるので、この最大公約数は

2.2.2 Subharmonic Summation に予想される欠点

スペクトル上のピークを求める方法は既に幾つか提案がある。しかし、Subharmonic Summation ではスペクトルからピッチ周波数らしさを求めるので、それらの手法の恩恵を受けることができない。

2.2.3 Multeapics 法

Subharmonic Summation では各ピーク (周波数 c_i, パワー p_i) それぞれが公約数らしさの分布を形成し、それらの和としてピッチ周波数らしさ $PL(x)$ が定義されていると捉えられる。本論文では $\sum_{n=1}^{N_h} h((c_i - \log n) - x)$ を下向き倍音列と呼ぶことにする。

このように Harmonic Summation を捉えてみると、$h(x)$ の性質のうち重要なのは、

条件1 中心で最大値をとり、急激に減少すること

であると考えられる。

さらに、公約数であれば複数のピーク周波数の値が分布にかかわってくる必要があると考えた。

そこで、それぞれのピーク同士の関係を分布に反映するために2つのピーク $(c_i, p_i), (c_i, p_j)$ を選び出し、すべての i, j の組合せにおいて最大公約数らしさをもとめそれを足し合わせれば、より精度を上げることができるのではないかと考えた。i, j から計算される最大公約数らしさの分布を $\mathrm{GCM}_{i,j}(x)$ と表すと、

$$PL(x) = \mathrm{GCM}_{1,1}(x) + \mathrm{GCM}_{1,2}(x) + \cdots + \mathrm{GCM}_{1,N_h}(x)$$

$$+\mathrm{GCM}_{2,3}(x) + \cdots + \mathrm{GCM}_{2,H_h}(x)$$

$$\cdots$$

$$+\mathrm{GCM}_{N_h-1,N_h}(x)$$

とおける。

もし、求めるピークが1つとわかっていれば、すべての下向き倍音列を足すのではなく掛けてしまえば $PL(x)$ のダイナミクスのオーダーが上がり、より安定して、ピッチ周波数を捉えられる予想される。しかし、対象としている楽曲は複数のピッチ周波数が存在するものなので、この手法では、実際のピッチ周波数の公約数は、非常に小さな値となってしまうと考えられる。

$PL(x)$ をもとめるのに2つのピークそれぞれの公約数らしさすべての和をとるのであれば、それぞれの公約数らしさ $\mathrm{GCM}_{i,j}(x)$ を求めるために、2つのピークが作る下向き倍音列の積をとっても問題ない。

また、2つのピークから求められる公約数らしさは p_i, p_j の関係から最大公約数として考えられない周波数帯においては、0 をとることが望ましい。よって $h(x)$ の条件1に加えて、

条件2 ある幅 2β 以上で0をとる

という条件を加えて作った下向き倍音列の積をとることで、2つの倍音列の両方が 0 より大きな値を持つ x でのみ値を持つような分布ができる。

さらに、公約数のうち最大ものを選び出すために、下向き倍音列を形成する各極大値の値に重みをかけ、低い周波数にいくに従い、小さい値になるように設定する。

n に対して単調減少な数列 $e(n)$ を考える。下向き倍音列を式で表せば、

$$g_i(x) = p_i \sum_{n=1}^{N_h} h((c_i - 1200 \log_2 n) - x) \tag{2.2}$$

ここで、N_h は考慮する倍音の数である、また $h(x; \alpha, \beta)$ は、ある α を中心に幅 2β の狭い範囲で正の値をもち、それ以外は0をとる関数である。この幅によって性質2が実現される。$e(j)$ は単調減少であるような数列で、$g_i(x)$ の包絡線を形成する。2つのピークについて $g_k, g_l (k \neq l)$ を乗算することにより、そのピークにおける最大公約数らしさを求める。g_k、g_l は N_h 個の狭い帯域以外では0をとるのでこれらを乗算をするとどちらか一方でも0を取る x での結果は0となる。2つの公約数を含む狭い範囲に対応する周波数帯に山を作る。こうして周波数上の二つのピークの公約数付近で極大値をとる分布となり性質1を満たす。N_p 個のピーク全てについて下向き倍音列中の2つの組合せで乗算を行いそれらを足し合わせピッチ周波数らしさとし、$PL(x)$ と書く。式で表せば、

$$PL(x) = g_1 g_2 + g_1 g_3 + g_1 g_4 + \cdots + g_1 g_{N_p}$$

$$+g_2 g_3 + g_2 g_4 + \cdots + g_2 g_{N_p}$$

$$+g_3 g_4 + \cdots + g_{N_p-1} g_{N_p}$$

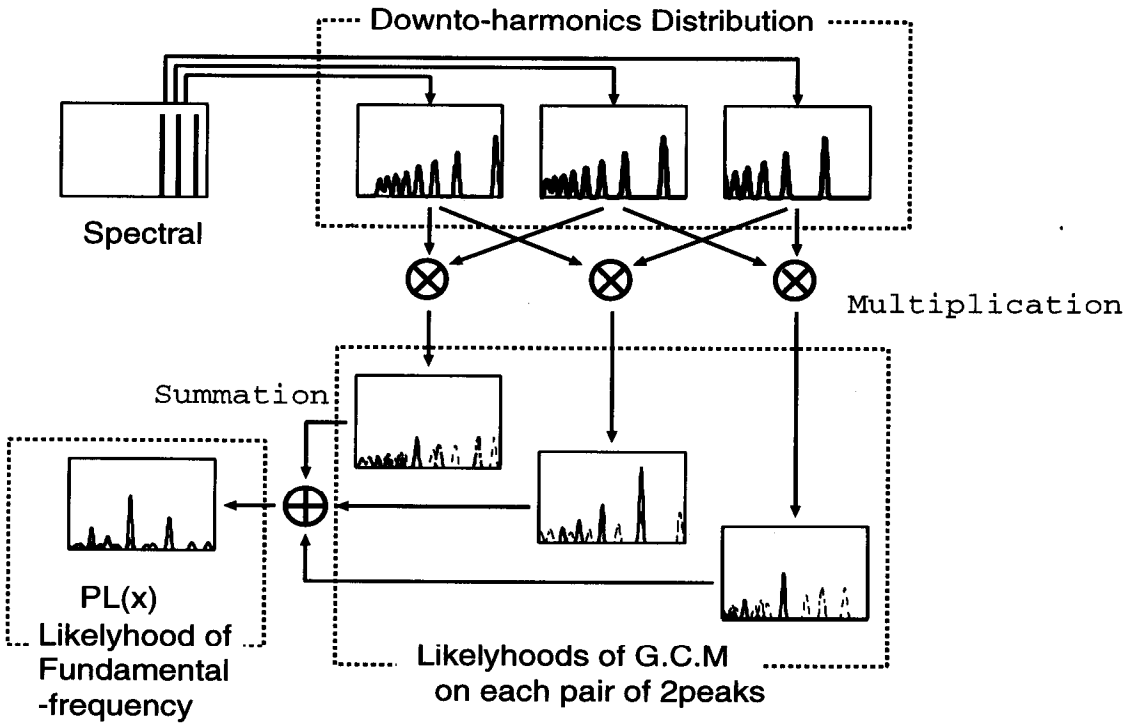

図 2.1: MULTEAPICS 法のブロック図

$$\cdots$$

$$+g_{N_p-1}g_{N_p} \tag{2.3}$$

となる。これを図示したのが、図 2.1 である。また、この式は

$$PL(x) = \frac{1}{2}\left\{\left\{\sum_{i=1}^{N_p} g_i(x)\right\}^2 - \sum_{i=1}^{N_p}\{g_i(x)\}^2\right\} \tag{2.4}$$

と簡単に書くことができる。この変形により計算量を減らすことができる。この手法を Multeapics 法と呼ぶ。

2.2.4 Multeapics の実装

Multeapics はさらに仮定を置くことにより、さらに計算がしやすくなる。式 (2.2) を式 (2.4) に代入すると、

$$PL(x) = \frac{1}{2}\left\{\left\{\sum_{i=1}^{N_p}\sum_{j=1}^{N_h} p_i e(j) h(x; c_i - \log j, \beta)\right\}^2 - \sum_{i=1}^{N_p}\left\{\sum_{j=1}^{N_h} p_i e(j) h(x; c_i - \log j, \beta)\right\}^2\right\} \tag{2.5}$$

となる。さらにここで

$$\log j + 1 - \log j > 2\beta \tag{2.6}$$

を満たす、という仮定をおく。これは、「1つの下向き倍音を構成する各 $h(x, c_i - \log j, \beta)$ は互いに重なることがなく、$g_i(x)$ に寄与する $h(x, \alpha, \beta)$ は x に対し高々1つである」、ということを表す。

すると、$a_1a_2 = a_1a_3 = a_2a_3 = 0$ のとき $(a_1 + a_2 + a_3)^2 = a_1^2 + a_2^2 + a_3^2$ となることを利用して、上式は

$$PL(x) = \frac{1}{2}\left\{\left\{\sum_{i=1}^{N_p}\sum_{j=1}^{N_h} p_i e(j) h(x; c_i - \log j, \beta)\right\}^2 - \sum_{i=1}^{N_p}\sum_{j=1}^{N_h}\left\{\{p_i e(j)\}^2\{h(x; c_i - \log j, \beta)\}^2\right\}\right\} \tag{2.7}$$

と表せる。$power_{iN_h+j} = p_i e(j), freq_{iN_h+j} = c_i - \log j, h_{sq} = \{h(x, \alpha, \beta)\}^2$ とおくと、最終的に $PL(x)$ は

$$PL(x) = \frac{1}{2}\left\{\left\{\sum_{l=1}^{N_hN_p} power_l h(x; freq_l, \beta)\right\}^2 - \sum_{l=1}^{N_hN_p}\left\{\{power_l\}^2 h_{sq}(x; freq_l, \beta)\right\}\right\} \tag{2.8}$$

と表せる。$h(x), h_{sq}(x)$ はテンプレートとして持って置くことが可能である。計算量もピークの数 N_h とテンプレートの幅 2β に比例することになりピークを求めるアルゴリズムとピーク数によっては、Subharmonic Summation よりも計算量が少なくなる。

2.3 提案する手法と Harmonic Summation との比較

1. スペクトルは漏洩によって、実際のピークを中心にスペクトルの山を作るが、Multeapics ではスペクトルの漏洩にかわる分布を意図的に付加する。
2. Subharmonic Summation ではスペクトルに対して処理を行うが、Multeapics では抽出されたピークをもとに計算を行う
3. Subharmonic Summation では、スペクトル上の1つの周波数におけるパワーについて計算を行うのに対し、Multeapics では2つの周波数におけるパワーを計算する。

2.4 Multeapics 法の利点

予想される利点を列挙する。

1. ピーク値の修正法を利用することができる
2. 漏洩のもつ多峰性に起因するピークの誤検出を防ぐ
3. 音源の種類に合わせて下向き倍音列をカスタマイズできる

2.5 評価

Multeapics 法の評価を行う。

2.5.1 評価方法

発音判断

$PL(x)$ は、ピッチ周波数らしさであり、この関数からある周波数 x が発音しているかどうかを判断する必要がある。本章では、SubharmonicSummation、Multeapics の両手法に対し、

1. 結果のリストをクリア
2. 周波数らしさを求める。
3. 周波数らしさのピークの最大値 $PL(x)$ とその値をとる周波数 x を求める
4. $PL(x)$ が閾値 P_{note_on} を超えていなければ終了
5. x の音がパワー $PL(x)$ で鳴っているものとして、結果のリストに加える。
6. $x - 50cent$ から $x + 50cent$ の範囲の倍音成分の帯域を 0 にする
7. 2. に戻る

という手順で発音判断を行った。閾値は SubHarmonicSummation、Multeapics それぞれについて $-70dB$,$-230dB$ を用いた。この閾値は予備実験によりおおよその値を求めて決定した。評価は

1. 評価用のデータの記述
2. SMF 化する
3. 2. で作成した SMF データから音響信号を合成する。(Timidity++を使用)
4. 2. で作成した SMF データから「時刻 t に $f$cent の音がベロシティ v で鳴っている」という情報を抽出する
5. Multeapics 法、HarmonicSummation 法の両手法を用い合成した音響信号から時刻 t に f cent の音が強さ p で鳴っている」という情報を抽出する
6. 4. と 5. を比較する。

という手順で行う。1. はテキストベースの独自のフォーマットで記述した。それを、SMF(Standard Midi File) に変換するプログラムを作成し、さらにソフトウェアシンセ Timidity++2.10.0 を用いて音響信号を生成する。

検出率の計算

検出率の計算方法を考える。検出したピッチ周波数を各鍵盤に対応する音の高さに割り当て、各鍵盤について、音が鳴っているまたは鳴っていないという 2 値を取るものとした。各時刻に対してシステムの結果と答えとで、全部で $2^{2^{88}}$ 通りの結果が存在する。つまり、

1. システムは検出せず、実際にも音がでていない
2. システムは検出せず、実際には音がでていた
3. システムは検出したが、実際には音がでていない
4. システムは検出して、実際にも音がでていた

の 4 通りのパターンが存在し、この 4 通りが時刻毎にピアノの鍵盤数 88 個だけ存在する。これらを、「.」「M」「F」「H」のシンボルを用いて画面に表示する。またすべての時刻でカウントしたものを N_0, N_1, N_2, N_3 と置き、

$$R_{recall} = 100 \times N_3/(N_1 + N_3)$$

$$
\begin{aligned}
R_{precision} &= 100 \times N_3/(N_2 + N_3) \\
R_{miss} &= 100 \times N_1/(N_1 + N_3) \\
R_{false} &= 100 \times N_2/(N_0 + N_2)
\end{aligned}
\tag{2.9}
$$

を計算した。

2.5.2 実験に用いたデータ

ソフトウェアシンセサイザ timidity++2.10.0 を用いて以下の 3 種類の音響信号を合成した。

1. 単音実験
2. 複音実験 1
3. 複音実験 2

それぞれの説明を行う。

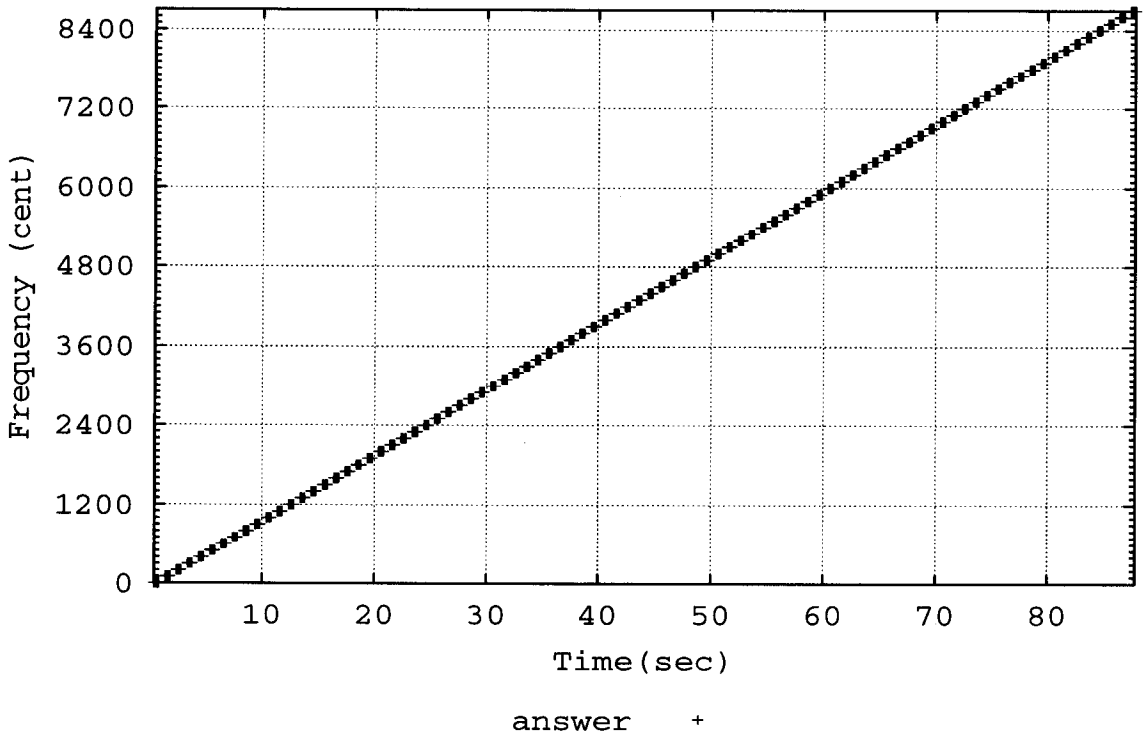

図 2.2: 単音実験に用いたデータ

単音実験 1

周波数による精度の違いを調べるため、テンポ 120(1 拍の長さは 500msec) で 1 拍音をならし 1 拍休むを繰り返す。ピアノの音色で A0(0) から C8(8800) まで半音ずつ順番に上がっ

てくるものを使用した。

複音実験 1

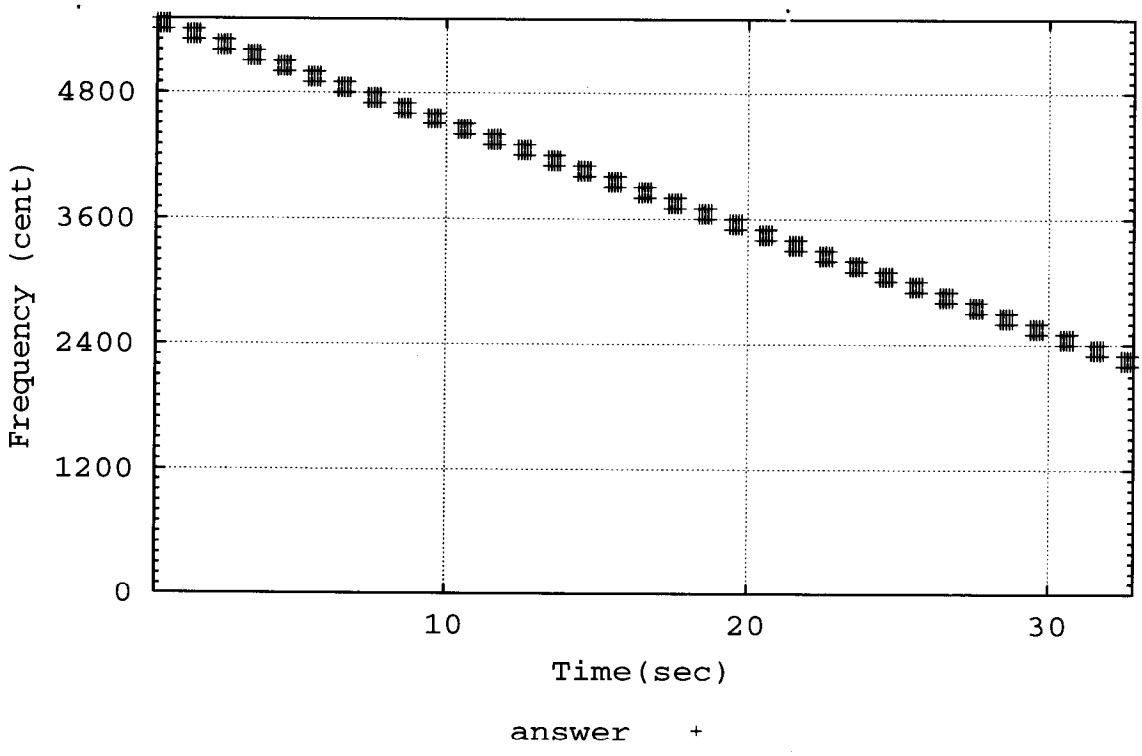

図 2.3: 複音実験 1 に用いたデータ

周波数の精度を確かめるため、半音 (100cent) だけ離れた高さの 2 つの音をテンポ 120(1 拍の長さは 500msec) で 1 拍音をならし 1 拍休むを繰り返す。ピアノの音色で D#5(5400cent)+E5(5500cent) の 2 つの音を同時に鳴らす。半音ずつ G2+G#2 まで音を下げていった。

複音実験 2

さらに音数が増えた時にどのような変化をするか調べるために、複音実験 1 で用いたものの低い音のさらに、400cent 低い音を追加し、3 音が同時に鳴っているデータを用いた。

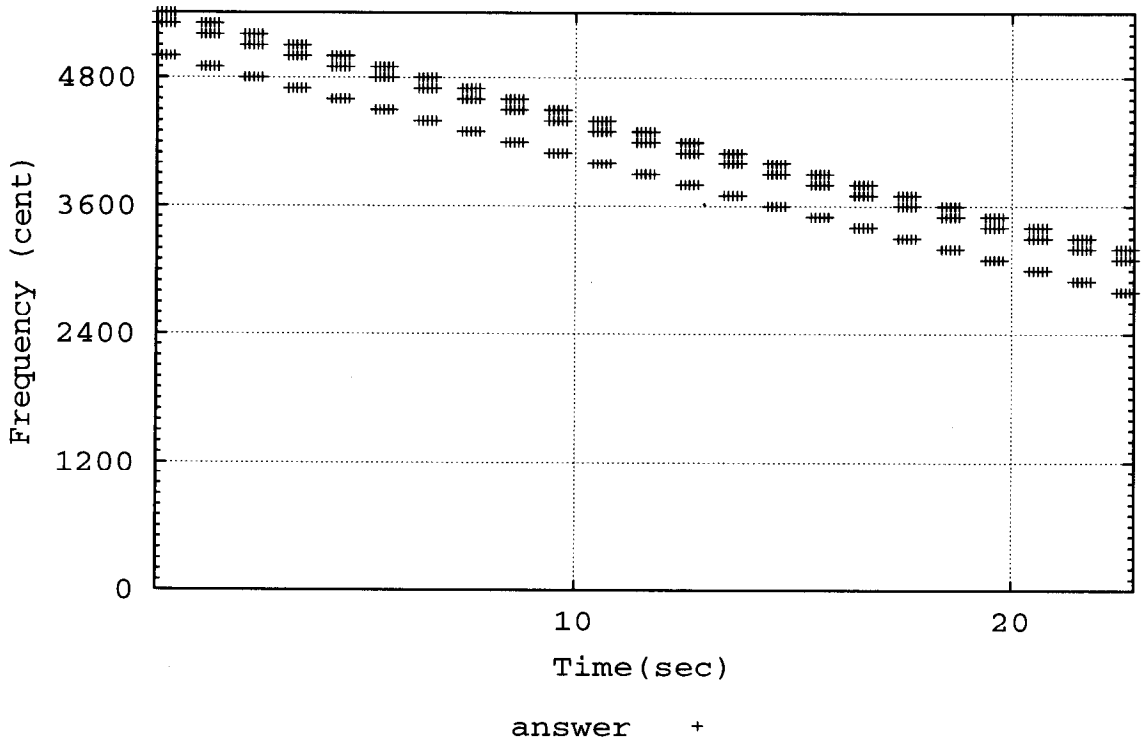

図 2.4: 複音実験 2 に用いたデータ

2.5.3 Subharmonic Summation による結果

まず Subharmonic Summation でどのくらいの精度がでるのか試す。周波数解析には 1024 点 FFT(ハニング窓) を用いた。

実験結果を表 2.1 に示す。結果はすべてにおいて成績が悪い。

それぞれの実験の中での特徴を見る。

単音実験

結果を図 2.5 に示す。単音実験では、全体的に false の数が多い。その中でも特にターゲットが F#4 以下である時間帯では A#1～A#2 付近の帯域にノートの発音とほぼ同じタイミングで実際には存在しないはずのノートを検出してしまっている。

ターゲット音が C7 以上の帯域では、音の立上り時に低い音を検出するものの、かなり良

表 2.1: FFT+Subharmonic Summation による結果

実験	R_{recall}	$R_{precision}$	R_{miss}	R_{false}
単音実験	47.7%	15.0%	52.3%	1.6%
複音実験 1	34.2%	16.9%	65.8%	1.9%
複音実験 2	1.7%	1.4%	98.3%	2.1%

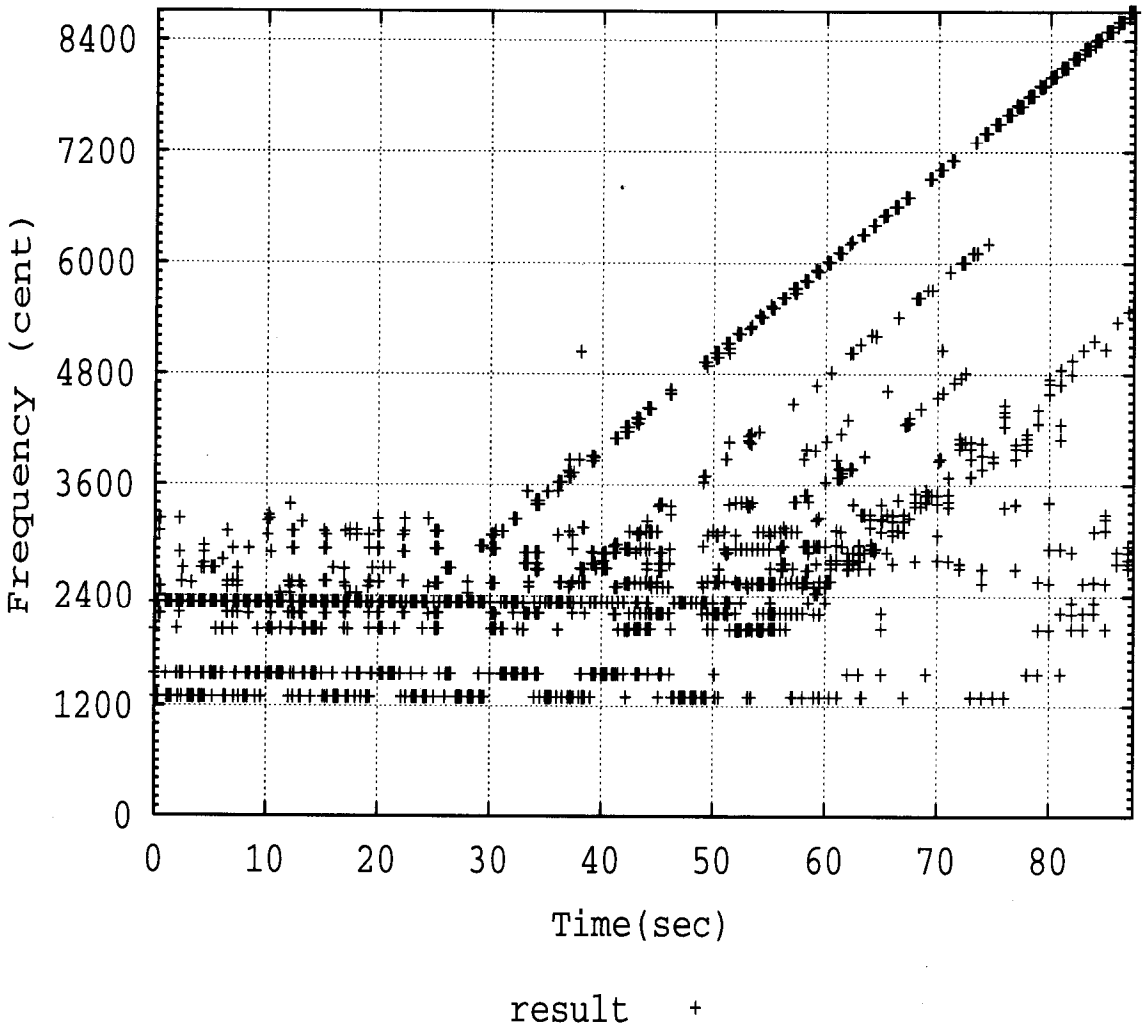

図 2.5: 単音実験の結果 (FFT+SubharmonicSummation)

い成績を出している。

ピッチ周波数らしさの分布を詳しく見てみる。

時刻 33400msec(ターゲット音は 3300cent) における分布を図2.6 に示す。ここでは、グラフを目で見る限りあきらかにピッチ周波数らしさを捉えている。しかし、検出されたピークは異なる位置であった。このピークの部分を拡大したのが、図 2.7 である。

これは、FFT によって、リニアの周波数上で計算された結果を対数の周波数で読むために、周波数が 10cent 動いても計算に使われる値が変わらず、ピーク判定条件の $p_{i-1} < p_i < p_{i+1}$ を満たさないため、と考えられる。リニアスケールからログスケールの変換を行う際に適切な補間を行う必要がある。

複音実験 1

結果を図 2.8 に示す。これらの分布 (図 2.9) から FFT によるスペクトルでは、2400 から 3600cent の帯域にかけて 2 つ存在すると思われるピークは混ざってしまい識別ができない。

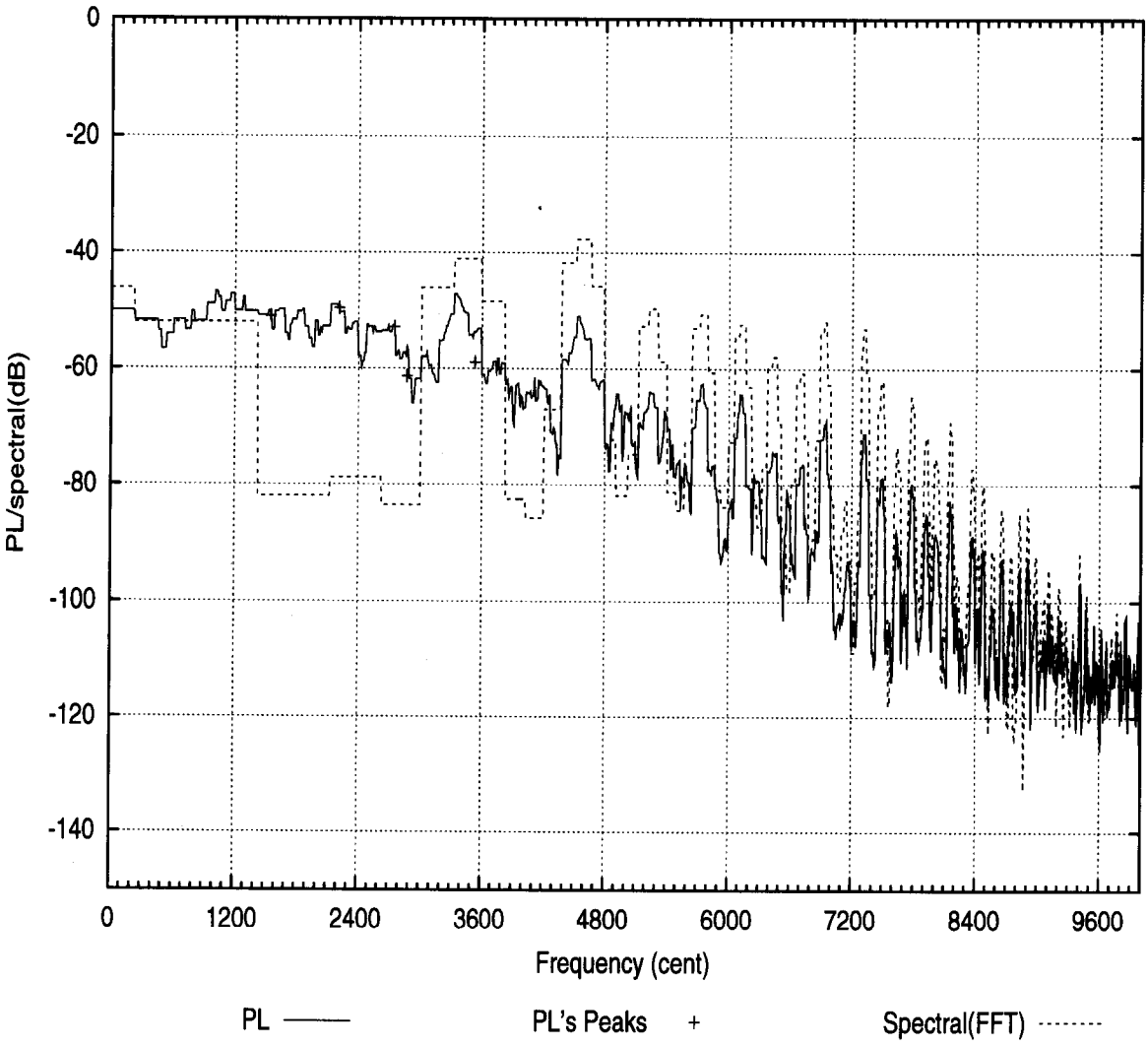

図 2.6: 単音実験 (33400msec)FFT+Subharmonic Summation

これに対しSubharmonic Summationの分布では、細かいピークが2つ存在しているのが分かる。これは、FFTが高い周波数に行く程周波数精度が上がって行く、という性質をうまく利用できると捉えることができる。

複音実験 2

結果を図2.10に示す。非常に精度が悪い。1300cent,1450cent,2050cent,2200cent,2300centで、元データとは、関係なく定常的にfalse alarmを検出している。

考察

全体的に非常に精度が悪い。ピッチ周波数らしさの分布を目で確認すると、正しい音域で極大値を取っているのにもかかわらず、

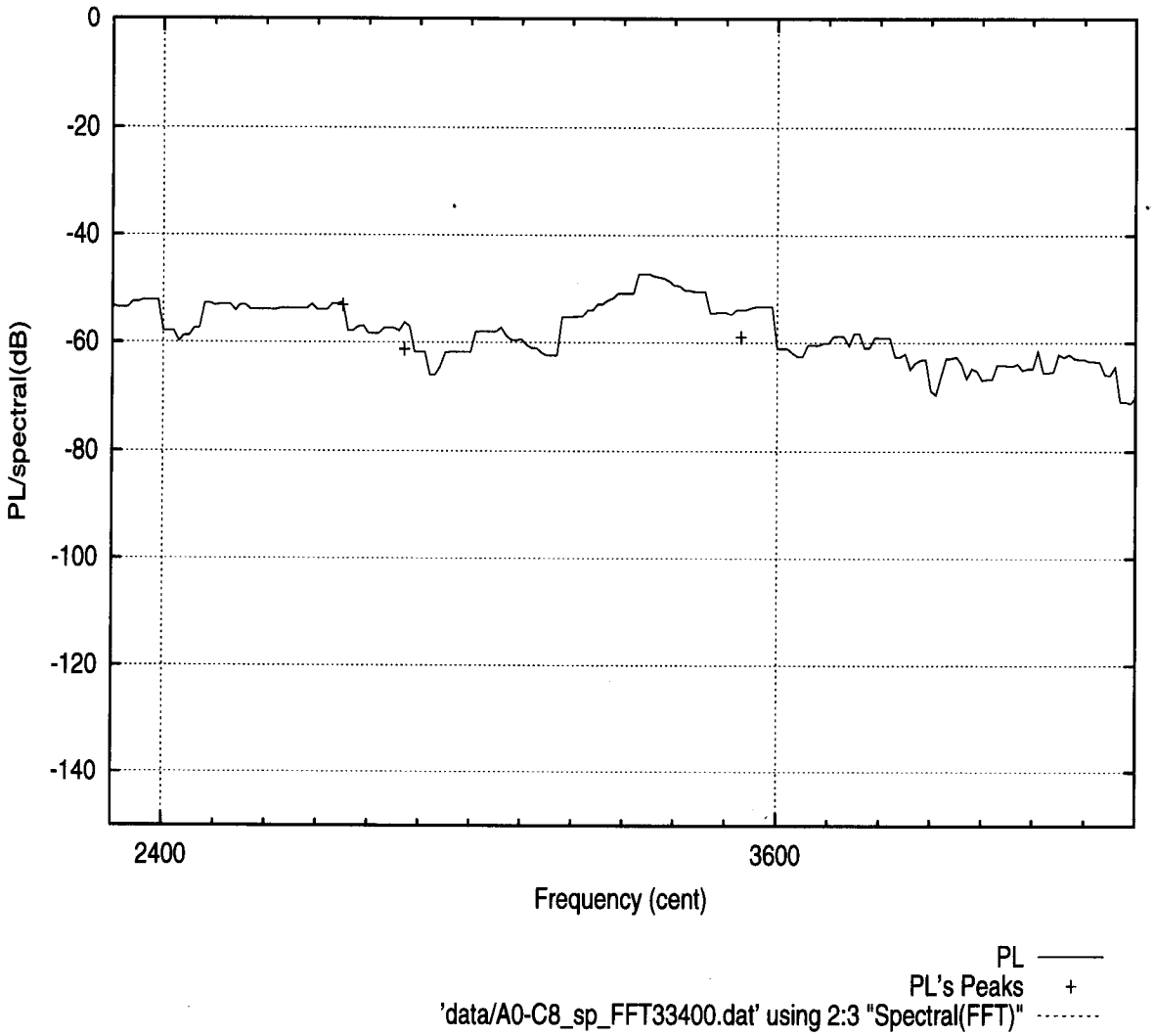

図 2.7: 2.6 を拡大したもの

1. 特にターゲットが低い音域の時に低音域のなだらかな山が検出されている
2. 極大値の値がしばらく続くため、ピーク検出条件から洩れている

という理由で正しい値を検出していない例が存在した。1. の原因については、周波数解析に FFT を用いているため周波数が低くなるにしたがって負の周波数成分からの漏洩の影響が強く現れるようになることが考えられる。

2. については、周波数解析の結果が周波数 (Hz) で等間隔で求まるのが原因であると思われる。結果が対数軸上に現れる解析方法を用いるのが良いと思われる。

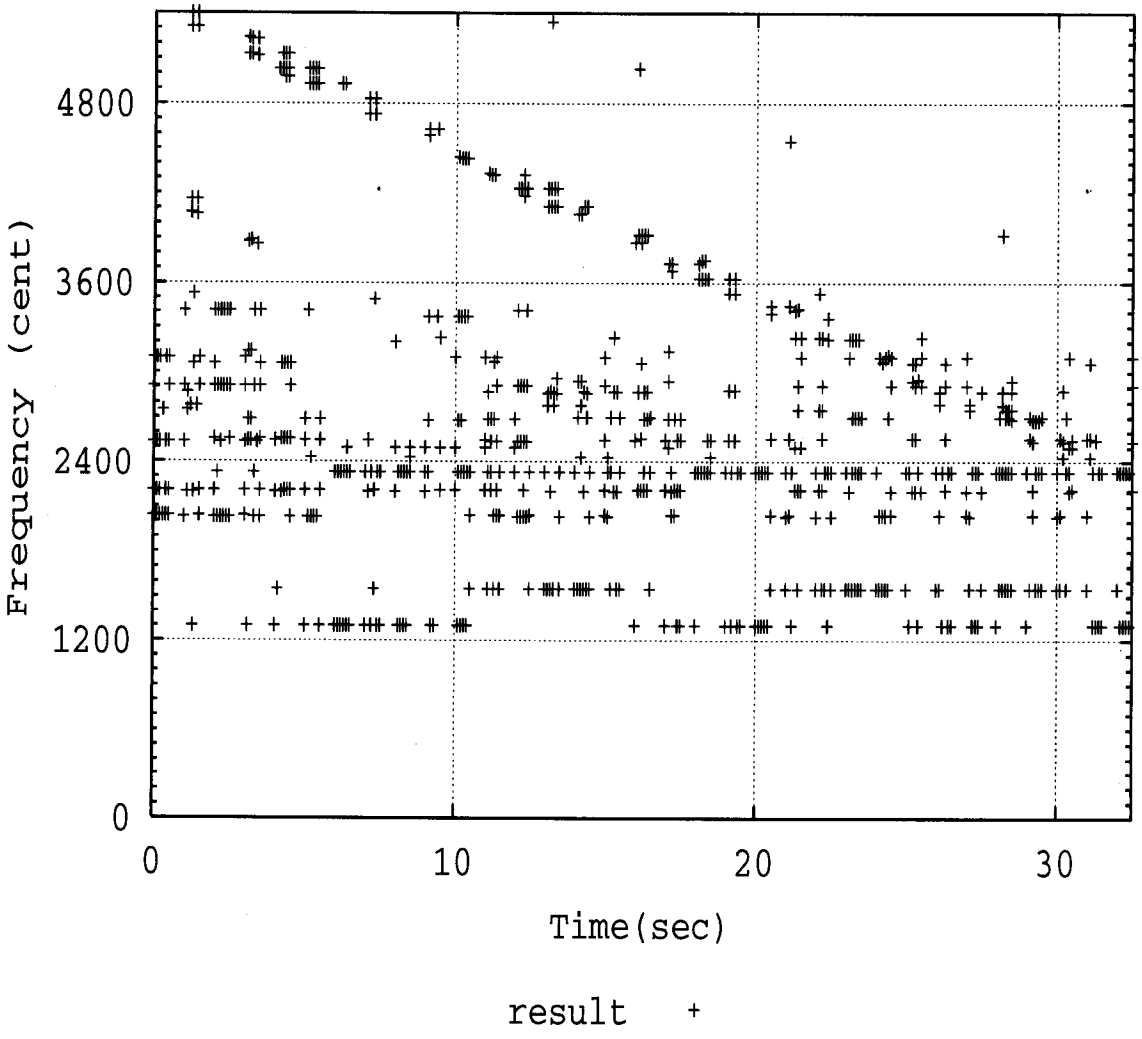

図 2.8: 複音実験 1 の結果 (FFT+SubharmonicSummation)

2.5.4 Harmonic Summation(可変窓長 DFT) による結果

前節では、周波数解析として FFT を用いたことが精度を落す原因として大きいと考えられる。

そこで、周波数解析に FFT ではなく、ハニング窓の窓長を解析する周波数によって変化

表 2.2: DFT(可変窓長)+Subharmonic Summation による結果

実験	R_{recall}	$R_{precision}$	R_{miss}	R_{false}
単音実験	37.5%	12.7%	62.5%	1.5%
複音実験 1	1.2%	0.7%	98.8%	2.8%
複音実験 2	38.5%	19.1%	61.5%	1.9%

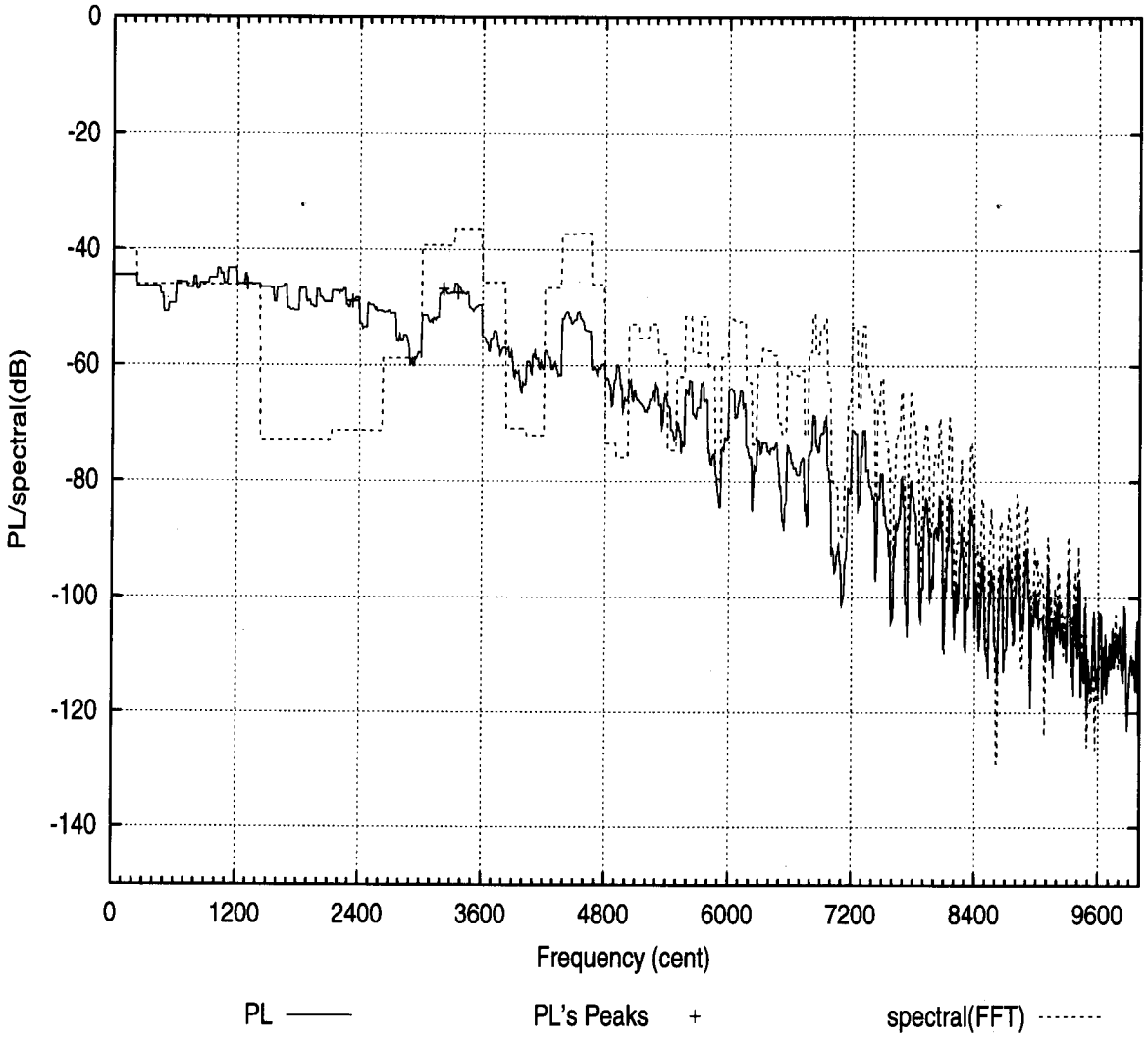

図 2.9: 複音実験 1(22400msec)FFT+SubHarmonicSummation

させ、常に解析区間内に同じ周期分だけ波が含まれる用に選んで測定を行う。式で表せば、

$$(W\phi f)(b,a) = \int_{-\infty}^{\infty} f(x)\overline{\phi\left(\frac{x-b}{a}\right)}dx$$

$$\phi(x) = \{0.5 + 0.5\cos(x/s)\}\exp(-\mathrm{j}x) \tag{2.10}$$

を用いた。なお実際の計算の際には正規化を行っている。解析区間内に含まれている波の数 s は 20 にした。各実験の結果を表 2.2 に示す。各実験について述べる。

単音実験

結果を図 2.11 に示す。精度が上がることが期待されたが、すべての率が悪くなっている。エラーの内容を見てみると、全帯域においてターゲットの 100cent 下の値を出力し、false alarm となっている例が非常に多い。

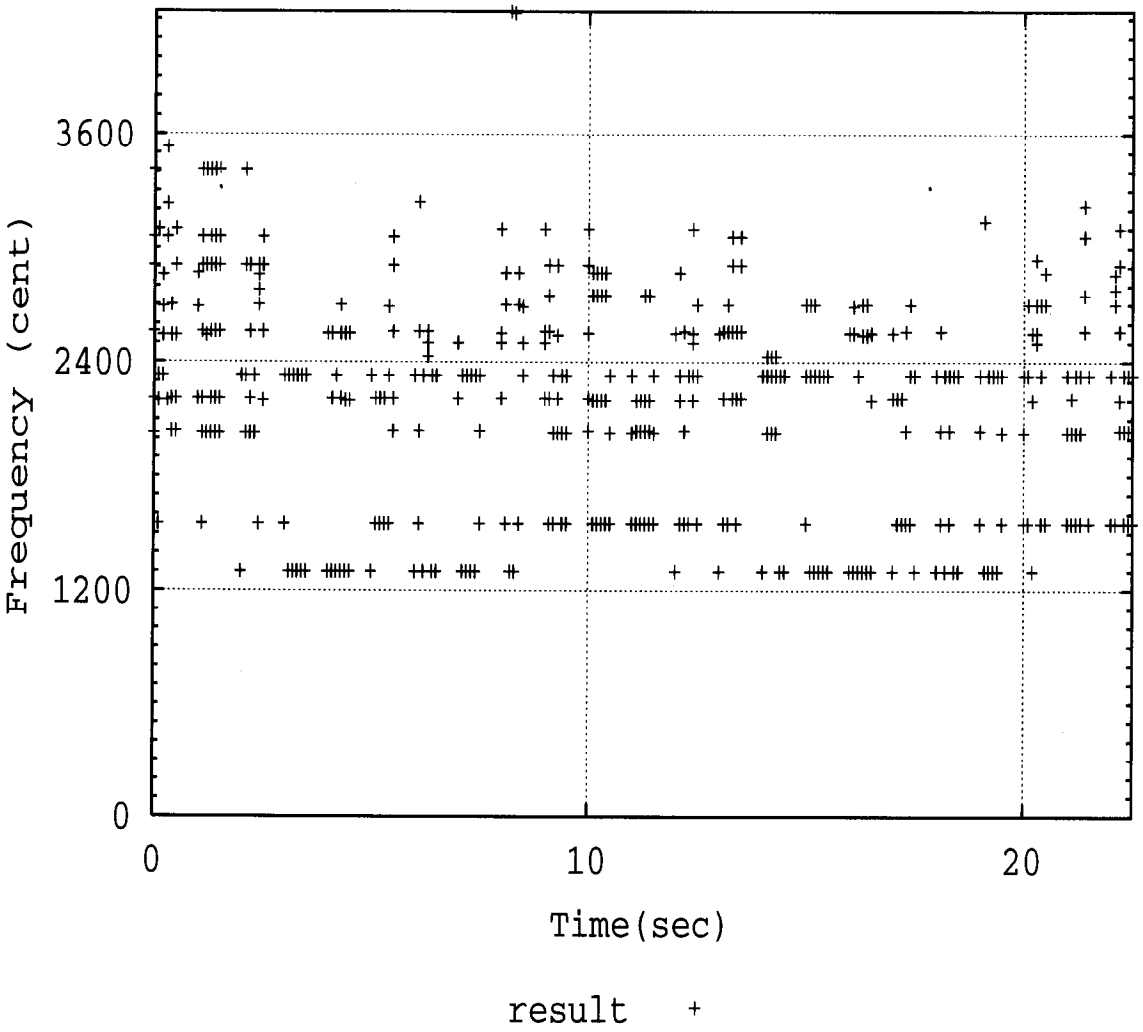

図 2.10: 複音実験 2 の結果 (FFT+SubharmonicSummation)

また、ピッチ周波数らしさの分布を見ると図 2.12 のように非常に狭い周波数帯域に多数のピークが検出されている。

これは音が出ているかを判断するアルゴリズム中で検出されたピークの ±50cent の成分を消しさった際に、新たなピークができ、それがピークとして検出される、ということでできていた。この問題を回避するためには、

1. 0 で埋める幅を長くし、新たにできるピークが閾値を超えないようにする。
2. wavelet の解析区間を長くとる

1. はスペクトル上の近接ピークが存在した場合となりのピークの山も削ることになり、近接音の認識精度下げる恐れがある。2. は時間精度が落ちること計算時間が増えることが欠点として挙げられる。

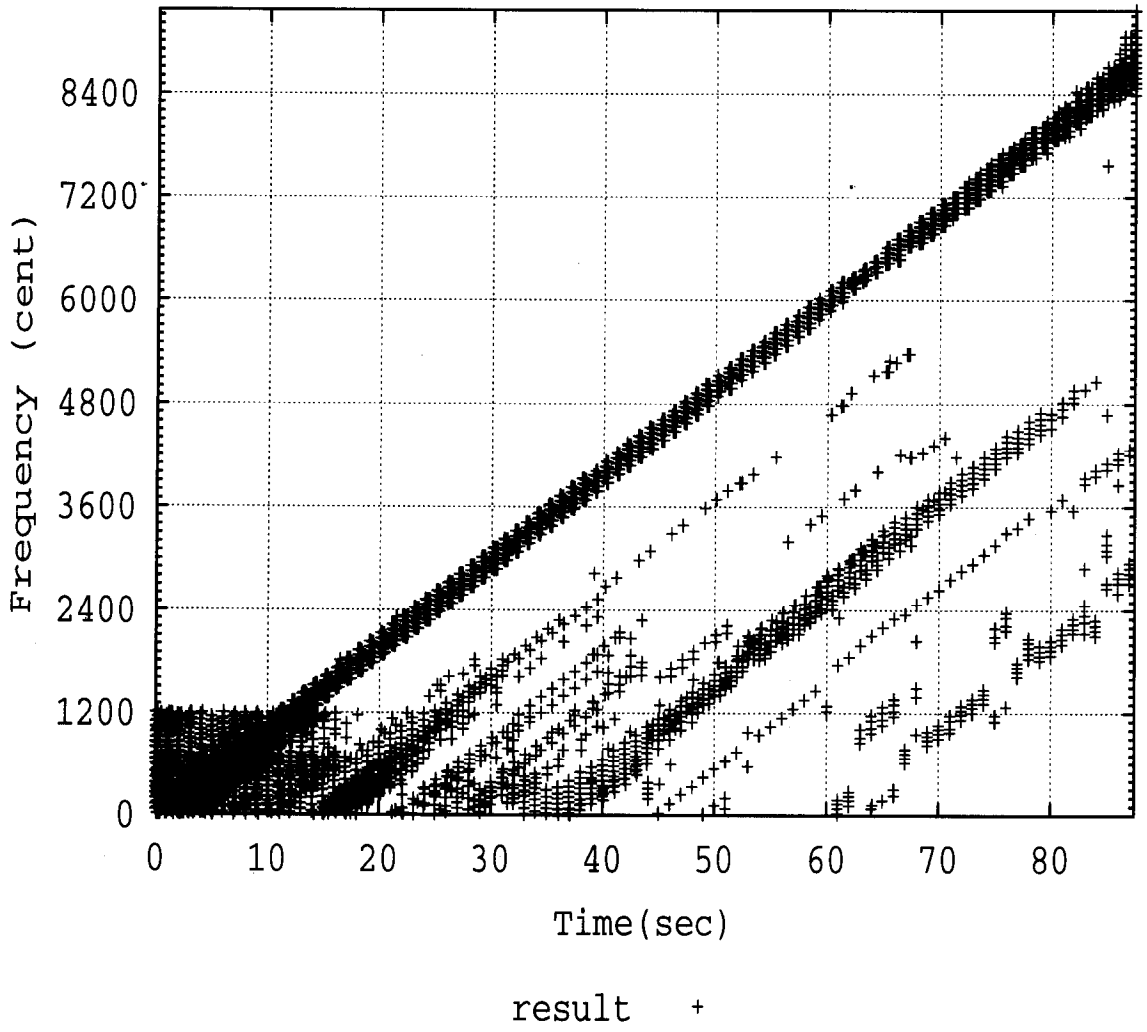

図 2.11: 単音実験の結果 (FFT(可変)+SubharmonicSummation)

複音実験 1

結果を図 2.13 に示す。これも精度が悪い。どのようなエラーかを見てみると、単音実験と同様に 1 つとなりの音を答えている例が多い。

分布を観察する。単音実験の時と同じように、狭い帯域に多数のピークが検出されている例があった。

複音実験 2

結果を図 2.14 に示す。

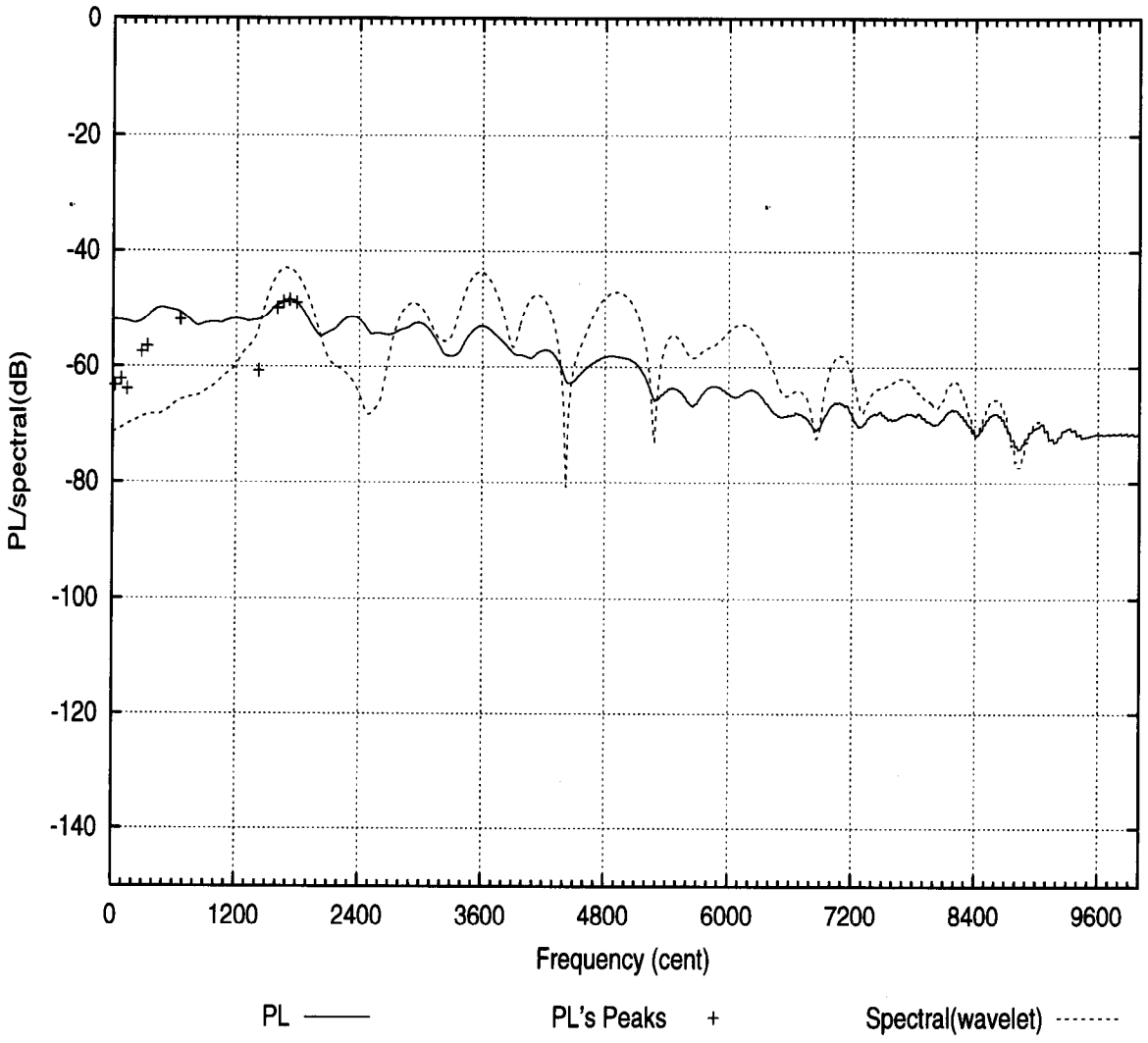

図 2.12: 単音実験 (17500msec) 可変窓長 DFT+SubharmonicSummation

2.5.5 Multeapics による結果

次に Multeapics の評価を行う。

Multeapics 法はスペクトル上のピークの列が分かった上で用いる手法なので、まずピークを求める。

表 2.3: FFT+Multeapics による結果

実験	R_{recall}	$R_{precision}$	R_{miss}	R_{false}
単音実験	81.4%	54.3%	18.6%	0.4%
複音実験 1	77.9%	67.6%	22.1%	0.4%
複音実験 2	38.6%	47.2%	61.4%	0.8%

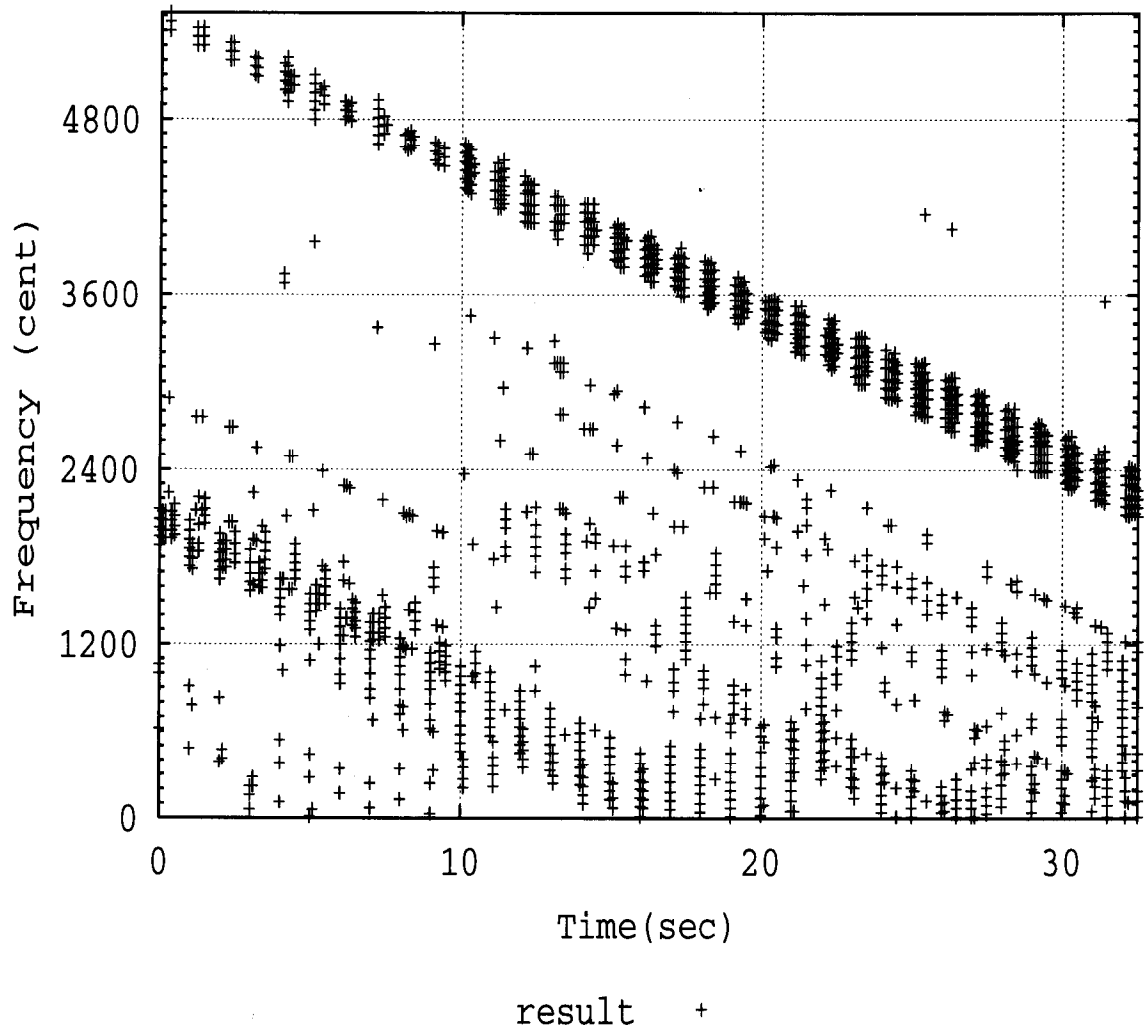

図 2.13: 複音実験 1 の結果 (FFT(可変)+SubharmonicSummation)

まず周波数解析には 1024 点 FFT(Hanning 窓) を用い、1 サンプルずらしてもう一度 FFT を計算し、それらの位相差から瞬時周波数を求める。あるピークが存在すれば、その付近の bin での位相の時間変化は等しくなる。よって、ある bin の中心周波数と瞬時周波数の差がなくなる周波数が実際のピークの周波数となる。

また $e(j)$ については、

$$e(j) = 1 - 0.05j$$

を用いた。

各実験による $R_{recall}, R_{precision}, R_{miss}, R_{false}$ を表 2.3 に示す。

単音実験

結果を図 2.15 に示す。エラーについて解析をすると、B1(1400cent) 以下の音程では、全くターゲットを捉えていない。しかし、C2 以上の音程ではほぼすべてのターゲットを捉えている。

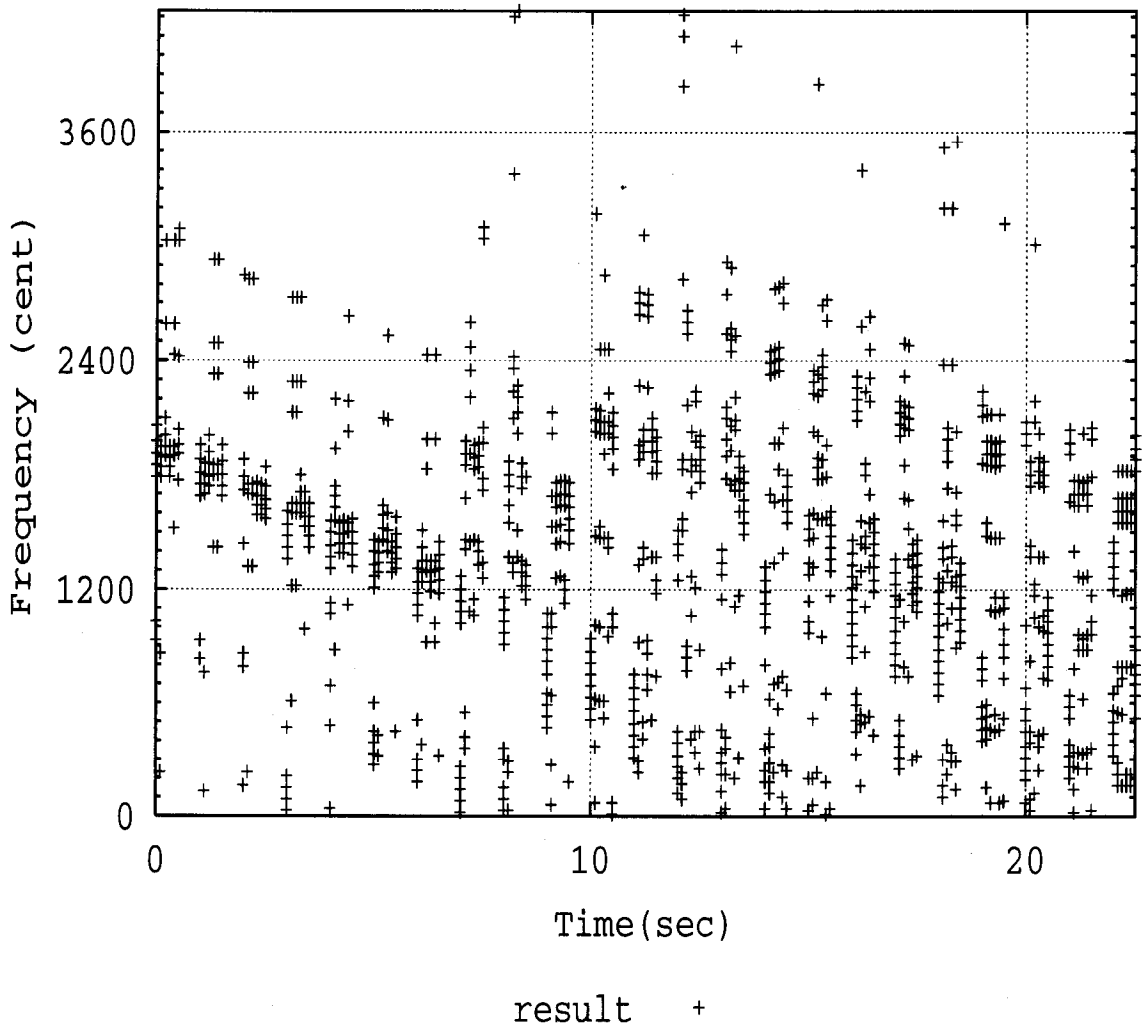

図 2.14: 複音実験 2 の結果 (FFT(可変)+SubharmonicSummation)

複音実験 1

結果を図 2.16 に示す。単音実験に比べて、少し miss が増えている。各時刻でのエラーを見ると、低い方のノートが C4 より高い音の高さにおいては、音の立上り時にすこしエラーはあるもののほぼ完璧に当てている。ここから徐々に false が増えてくるものの、低い方の音が E3 くらいまでの帯域では false が出現する周波数がターゲット周波数に依存したものではないと思われる。よって、「同じ音がある一定以上の時間鳴っていないければ、ノイズと思って無視する」等の規則を付加することで、この帯域までは、識別ができると思われる。

E3 以下になると、2 つの音のうち 1 つしか検出できなくなってくる。

E3+F3 は、LIL(Low Interval Limit:2 つのノートから構成される各音程毎に定められた音の高さで、コードとしてのサウンドの明瞭さを失わない最低音を表している) における短 2 度の最低音である。この音域で識別ができているので、2 音の近接音の周波数精度については問題がないと言える。

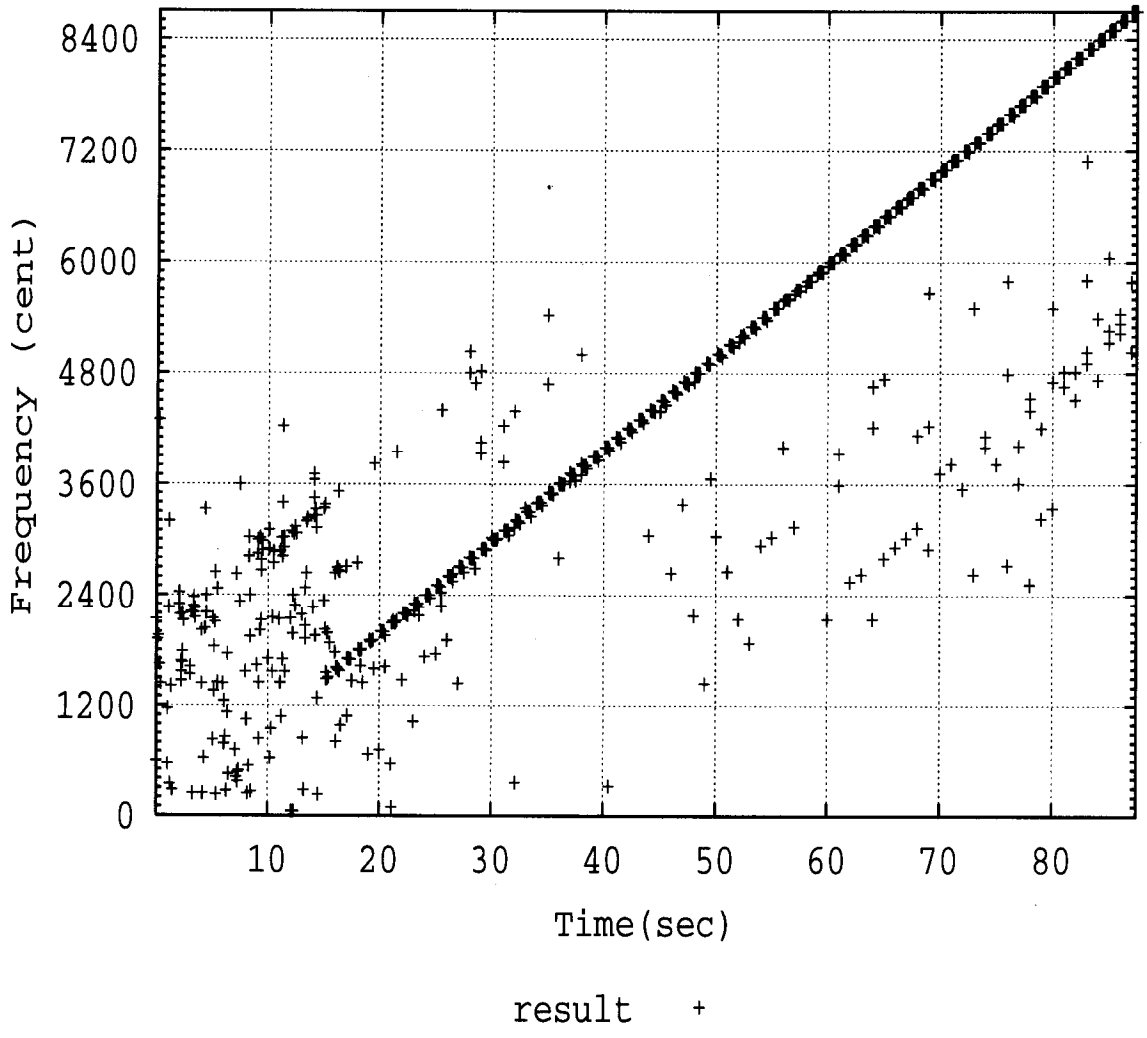

図 2.15: 単音実験の結果 (FFT+Multeapics)

複音実験 2

結果を図 2.17 に示す。先の 2 つの実験に対してかなり精度が落ちている。各時刻毎のエラーを調べてみると、500msec 鳴っているターゲットのすべてを当てている帯域は存在せず、最高でも 300msec～400msec 程度であった。さらに、false が起こる周波数についてはターゲットの周波数に対して、一定の音程を置いた位置でエラーが検出されており、これは時間的な連続性を用いて束縛をかけても除去できないエラーとなると考えられる。

開始から 800msec 分の結果を抜きだしたものが図 2.18 である。

下方向は時刻を表しており、1 行が 100msec を表す。右方向にノートナンバーを示している。最上段にはノートナンバーに対するスケールを示しており、書いてある数字は、オクターブ数を表している。各キャラクタは

1. 「.」: ターゲットも鳴っておらず、システムもしていない
2. 「M」: ターゲットは鳴っていたが、システムは検出していない
3. 「F」: ターゲットは鳴っていなかったが、システムは検出した

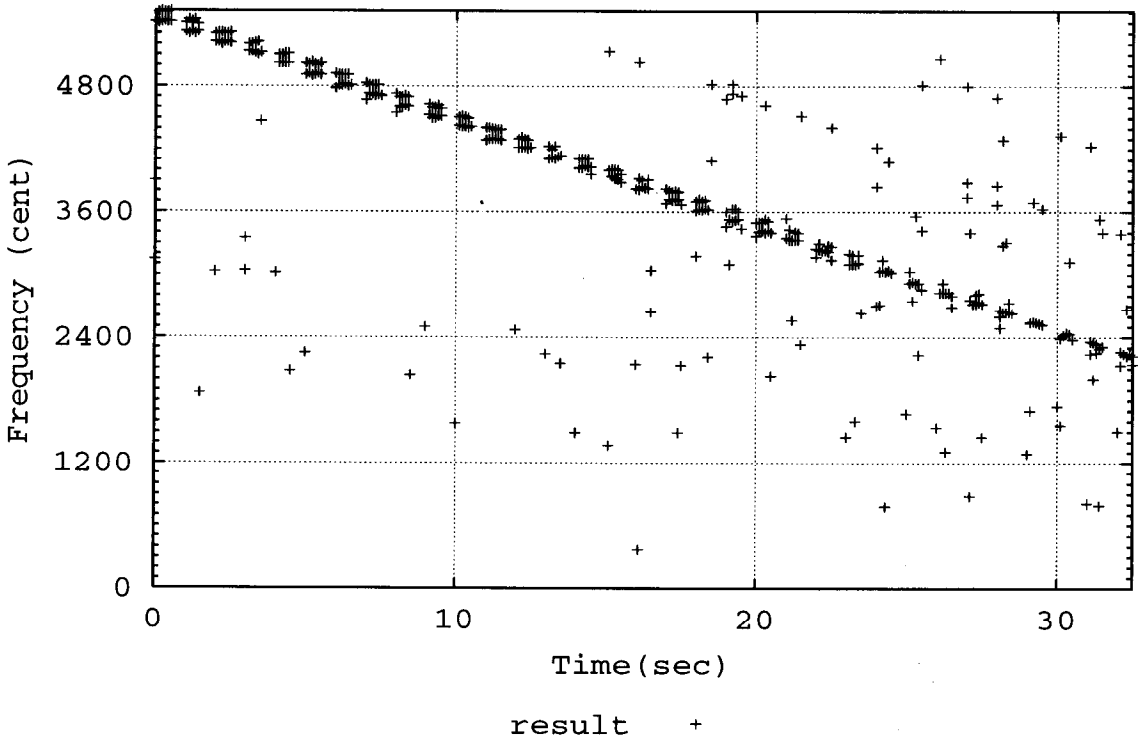

図 2.16: 複音実験 1 の結果 (FFT+Multeapics)

4. 「H」: ターゲットは鳴っていて、システムも検出した

を意味する。例えば図 2.18 の場合、400msec の時点では全く音が鳴っていないにもかかわらず、D#3 と B4 を検出したことになる。

この結果を見ると、300,500msec では 3 音すべてを当てているが、300,500,600msec では、低い音域を検出してしまっている。特に 300,500 のような、ターゲット音の一番低い音に対して、2 オクターブと長 3 度低いノートを誤検出する例が約 7000msec までの時刻で多く認められた。このようなエラーの場合、必ずターゲットのノートは 3 音とも miss していた。これは低い音が誤検出去れその倍音成分の除去の際に、ターゲット音の主要な倍音成分が除去されてしまったものと考えられる。

200〜500msec 間の 4 時刻のピッチ周波数らしさの分布をそれぞれ図 2.19 ,2.20 ,2.21 ,2.22 に示す。分布は、各時刻にノートを決定する際にピーク列を求めた一番始めに用いた分布と、その時刻に選ばれた音を表したものである。図 2.19 ,2.21 からは、分布から求められる最大のピークのみを用い、繰り返して分布を求めることで、低音域に出ている余計な成分がうまく除去されていること、閾値を低めに設定しても、期待しないピークが多く検出されることがないことが分かる。逆に、図 2.20 ,2.22 からは、最初の分布の計算の時点で最も強い成分を低域に取られてしまうと、大きな誤差が出ることが分かる。低域の誤検出は $e(j)$ の値を検討することで減らすことができると思われる。

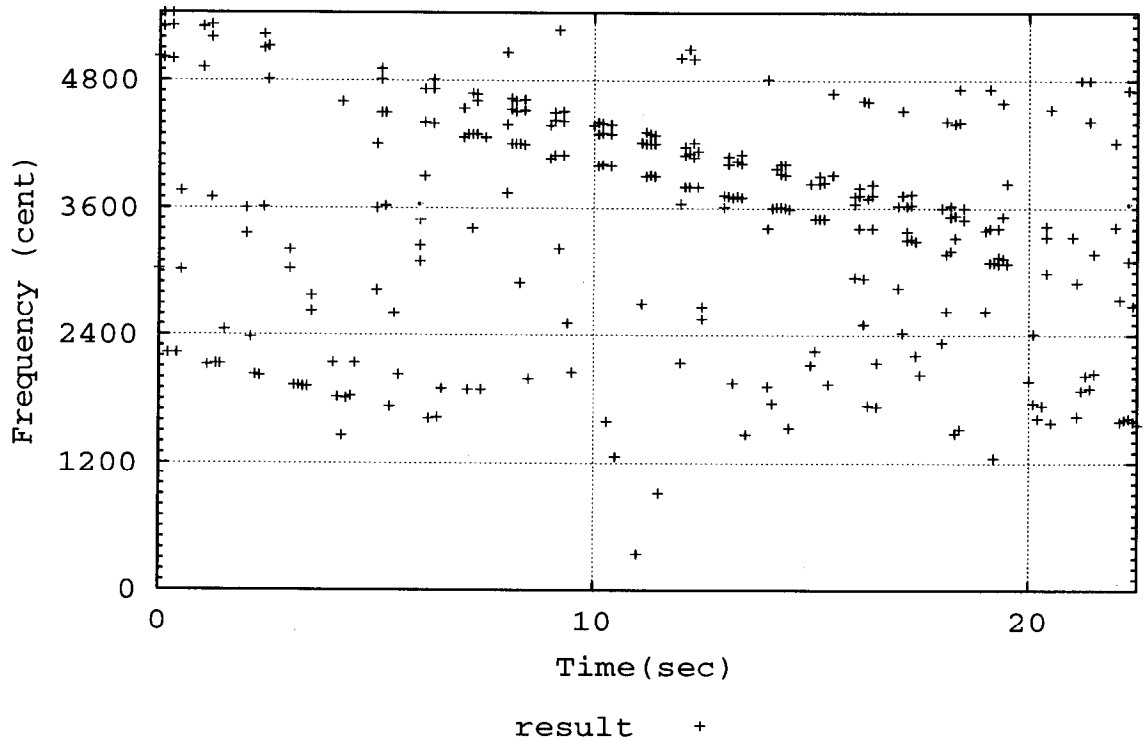

図 2.17: 複音実験 2 の結果 (FFT+Multeapics)

2.5.6 Multeapics(可変窓長 DFT) による結果

Multeapics 法を可変長 DFT でも試してみる。
各実験による $R_{recall}, R_{precision}, R_{miss}, R_{false}$ を表 2.4 に示す。

```
2----------3----------4----------5----------
..........................................
..............F..................F..........
...................................H..HH.....
.......F...........................M..MM.....
...................................H..HH.....
.......F...........................M..MM.....
..............F.......F...........M..MM.....
..........................................
```

図 2.18: 複音実験 2 の結果 (開始から 800msec まで) 周波数解析:1024 点 FFT(瞬時周波数を用いてピークを抽出) ピッチ推定法:Multeapics

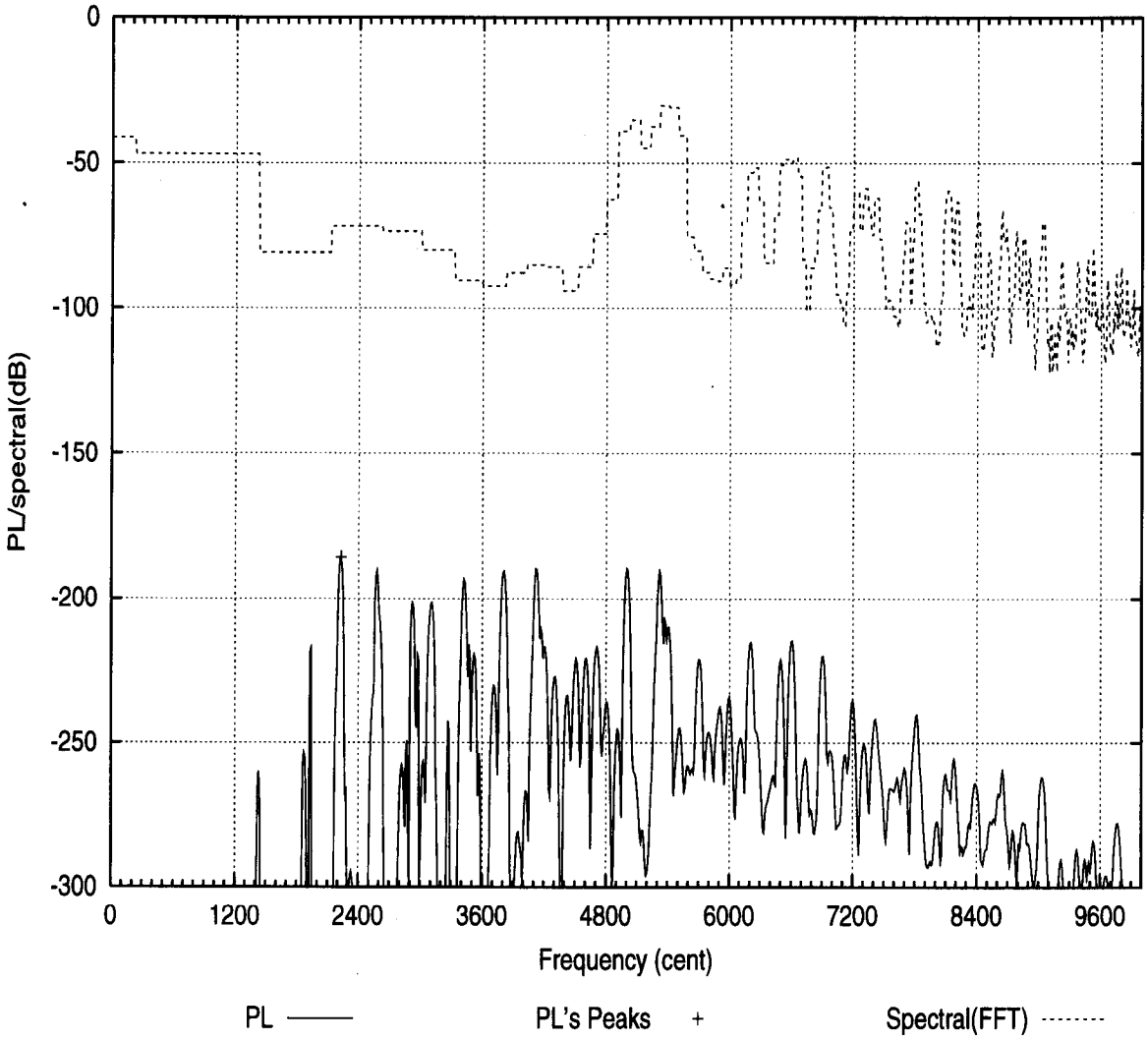

図 2.19: 複音実験 1(200msec)FFT+Multeapics

表 2.4: DFT(可変窓長)+Multeapics による結果

実験	R_{recall}	$R_{precision}$	R_{miss}	R_{false}
単音実験	95.2%	46.0%	4.8%	0.6%
複音実験 1	77.6%	53.3%	22.4%	0.8%
複音実験 2	43.8%	41.0%	56.2%	1.1%

単音実験

結果を図 2.23 に示す。$R_{recall} = 95.2\%$ と非常に良い成績が出ている。しかし、$R_{precision}$ が低い。これは特に始めの 15000msec 間で C2 以下の低音にたくさんの false があるのが原因と考えられる。

時刻毎に見ると、最初の A0(=0cent) の音がすべて miss したがそれ以外のターゲットは

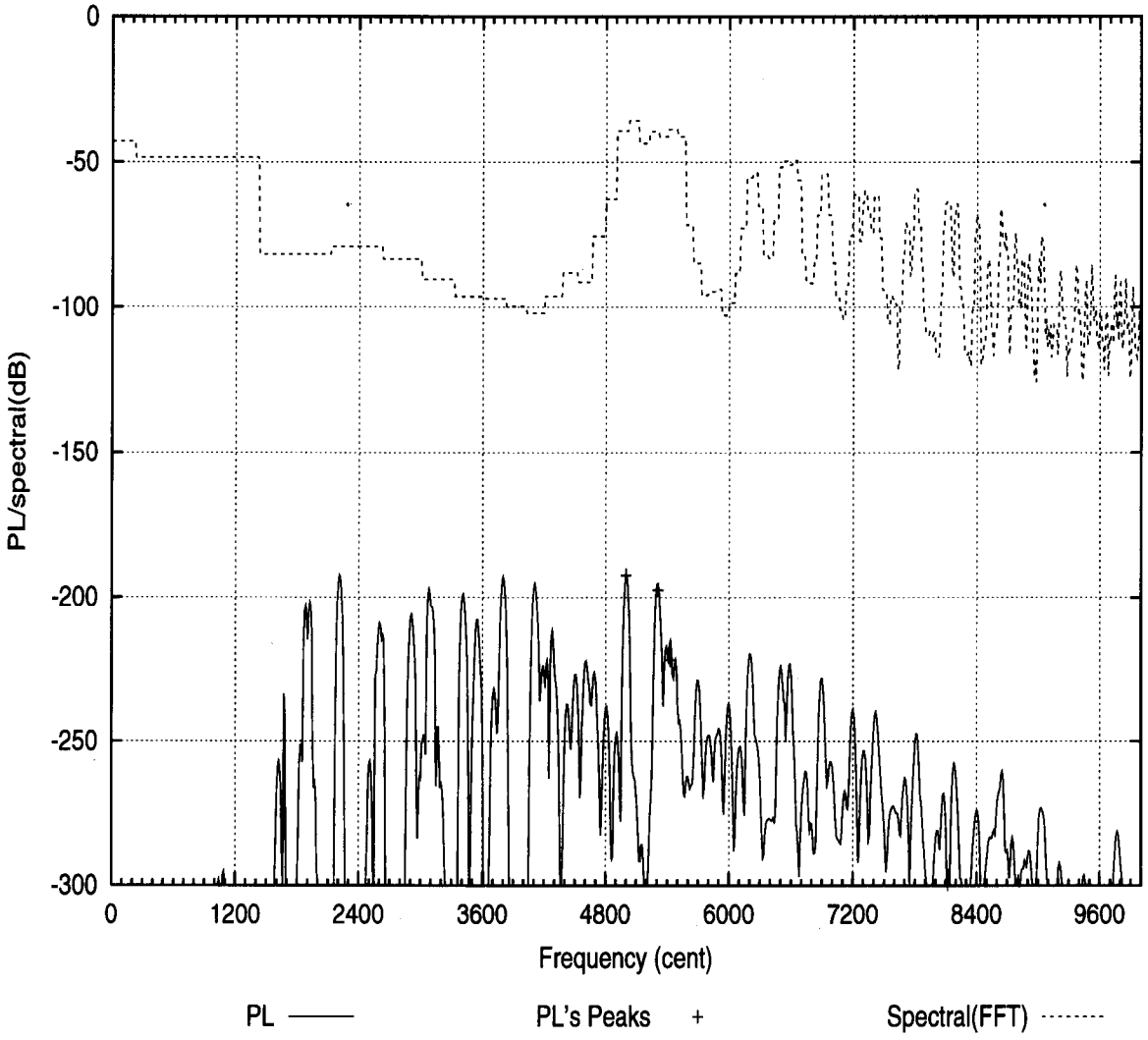

図 2.20: 複音実験 1(300msec)FFT+Multeapics

ほぼ 100%捉えている。false の周波数ははじめの 700msec については同じ周波数が隣り合う時刻で検出されているものが多いが、それ以降は定常的なものではない。よって時間連続性を導入することで、かなりの精度向上が見込める。

最初の A0 を失敗したのは、分布の計算が 0cent から始まっているので、ピーク条件を満たしようがなかったからだと思われる。

時刻 300msec における分布を 図 2.24 に示す。低域を拡大したものが図 2.25 であるが、明らかに 周波数 0 におけるピーク成分があるにもかかわらず、ピーク判定条件を満たさないめに検出されていない様子がわかる。そして、その 1 オクターブ上の音程が検出されてしまっている。これは、低い周波数用に解析用バッファにあと 10cent 足せば、A0 における miss だけでなく、大量の false もなくすことができると考えられる。

これらを考慮すると、単音検出ではほぼ 100% の精度が期待できる。

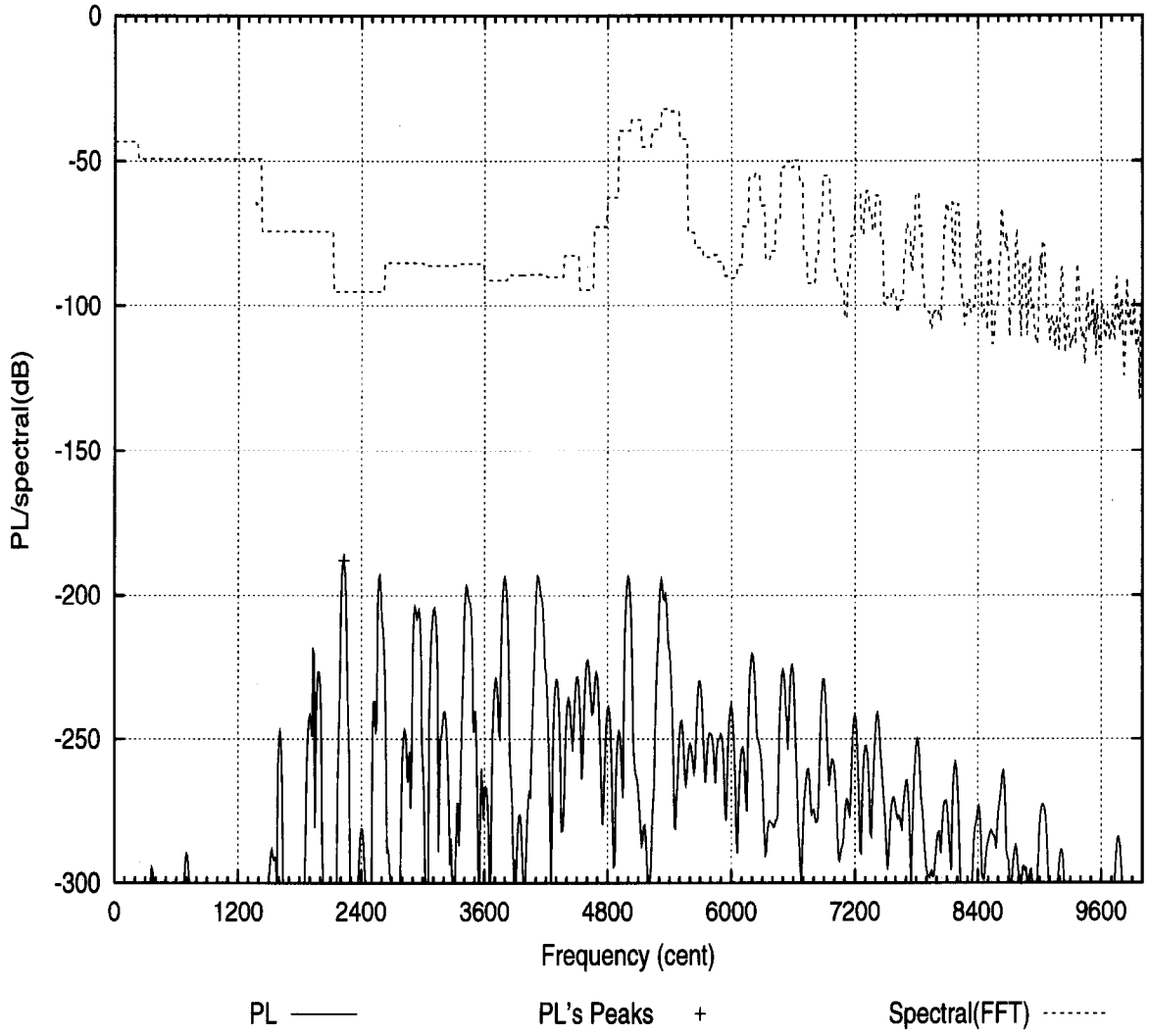

図 2.21: 複音実験1(400msec)FFT+Multeapics

複音実験 1

結果を図2.26に示す。FFTに対して精度向上が見込まれたが、結果は精度が落ちている。

分布のグラフを見ると、半音間隔で2つのピークを検出して欲しいが、山が重なって一つになってしまっているのが分かる。

FFTの時は、高い周波数帯でターゲット音の高調波成分と思われる2つのピークがはっきりと確認できる (図2.27) が、可変窓長DFTを用いた時はそれがなだらかに観測されていた。

複音実験 2

結果を図2.28に示す。ターゲットの一番低い音の2オクターブ下のノートが誤検出される例が多い。これはターゲット音の一番下と2番目の倍音成分すべてがターゲットの一番低い音の2オクターブ下の音の倍音成分の周波数に対応するためと思われる。

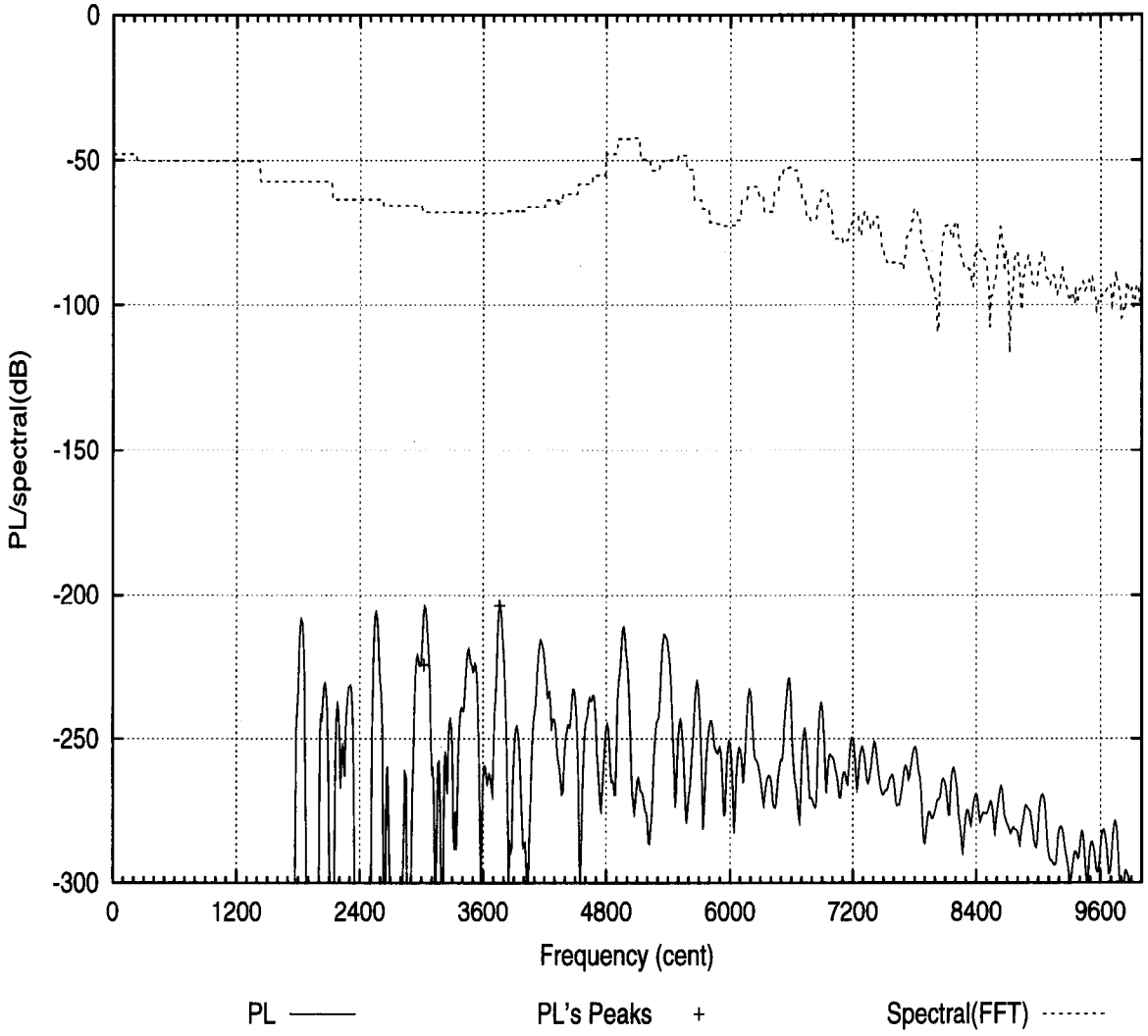

図 2.22: 複音実験1(500msec)FFT+Multeapics

ターゲットの一番下の音と2番目の音の差は長3度であるが、これを周波数(リニア)比で表すと約4:5である。これらの最大公約数の周波数は1となり、ちょうど2オクターブ下の音の高さに当たる。

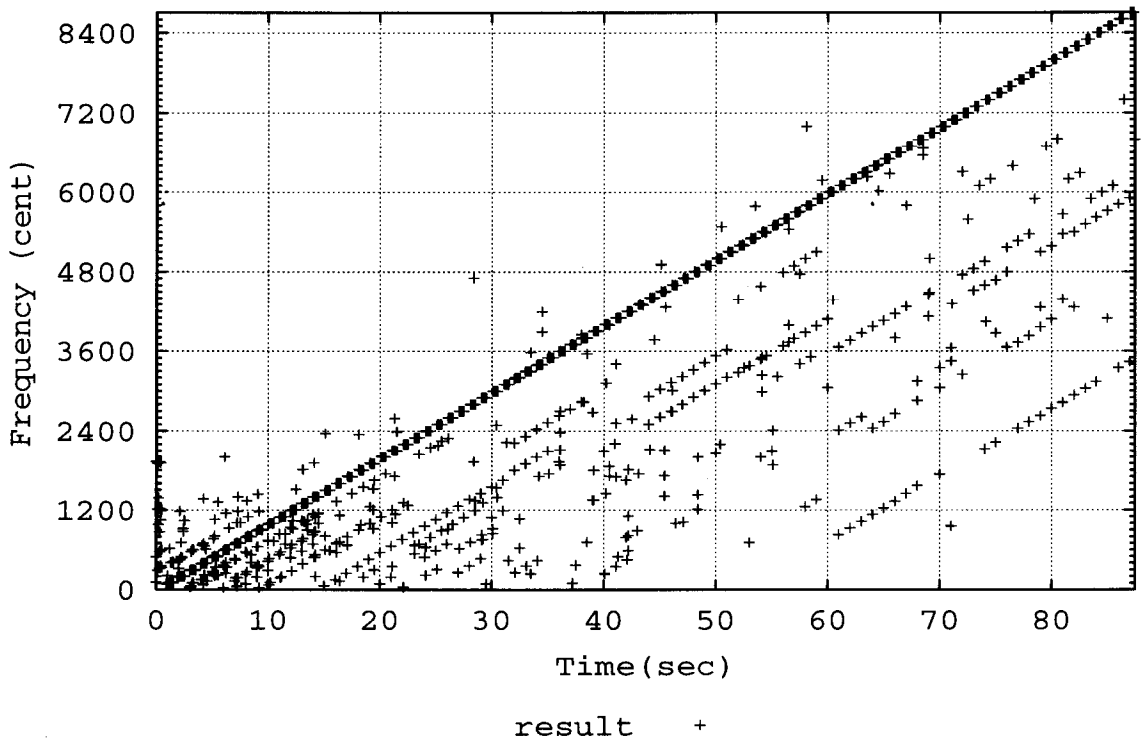

図 2.23: 単音実験の結果 (可変長 DFT+Multeapics)

2.6 考察

本実験では、すべての実験で Multeapics 法の精度が、Subharmonic Summation の精度を上回る結果がでた。しかし、Subharmonic Summation のエラーは前処理によって除去できるものが多く、この手法の正当な結果は得られていない。

これらを検討しもう一度評価を行う必要がある。

しかし、実験結果から Multeapics 法の利点として、

1. ピークの集合からピッチ周波数らしさを求めるのでピーク値の補正法を使える
2. 周波数解析に用いる手法を意識した前処理を行う必要がない

を挙げることはできる。

2.7 今後の課題

本実験だけでは Multeapics 法の評価をするには不十分である。今後は、

1. 対雑音性がどの程度とれるのか
2. 2つピークすべての組合せについて下向き倍音列の積を取ることの効果
3. $e(j)$ として用いる関数には何が良いか
4. より楽音に近い対象 (打楽器が入っている等) に適用した時の性能

について評価、検討を重ねていく必要がある。

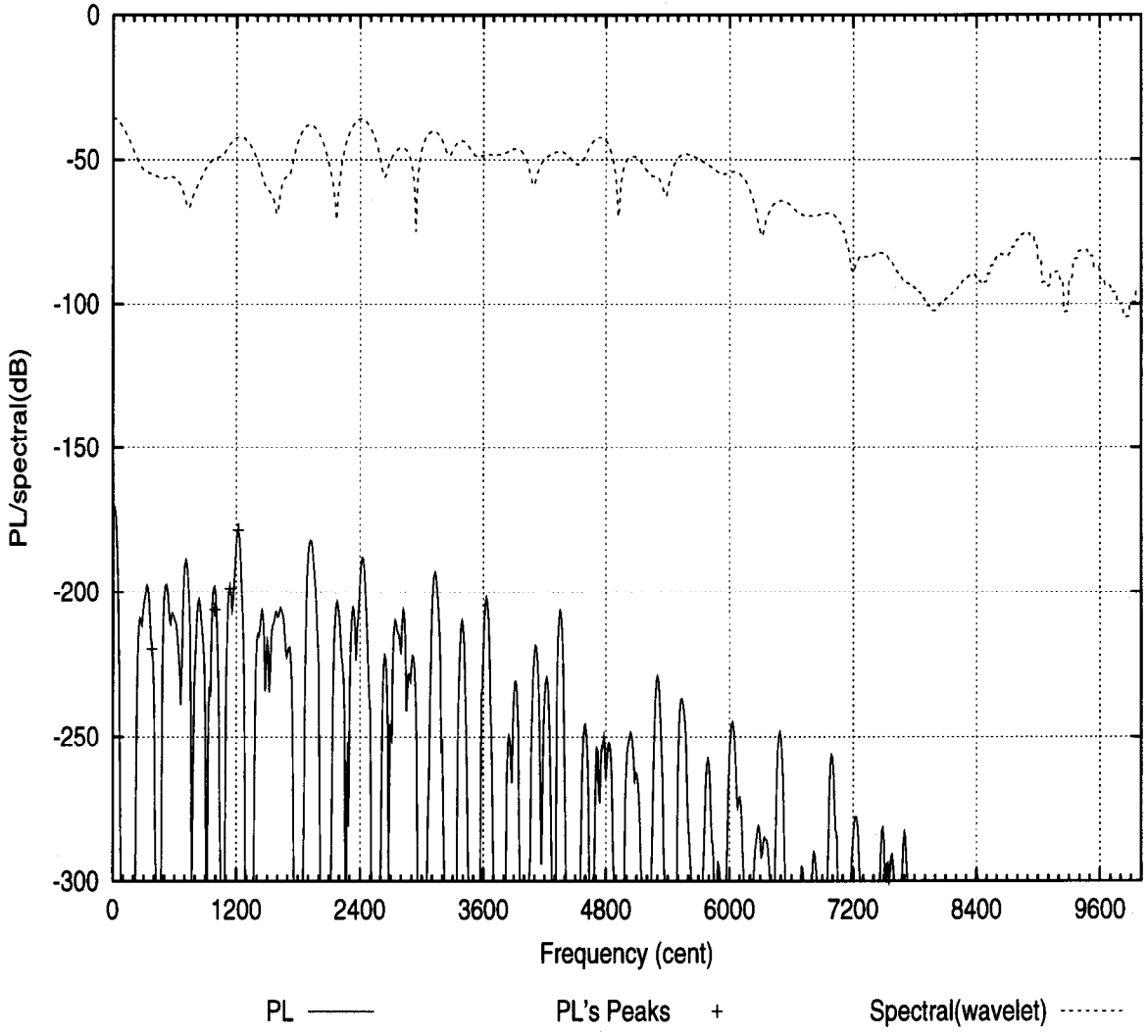

図 2.24: 単音実験 (300msec)FFT+wavelet

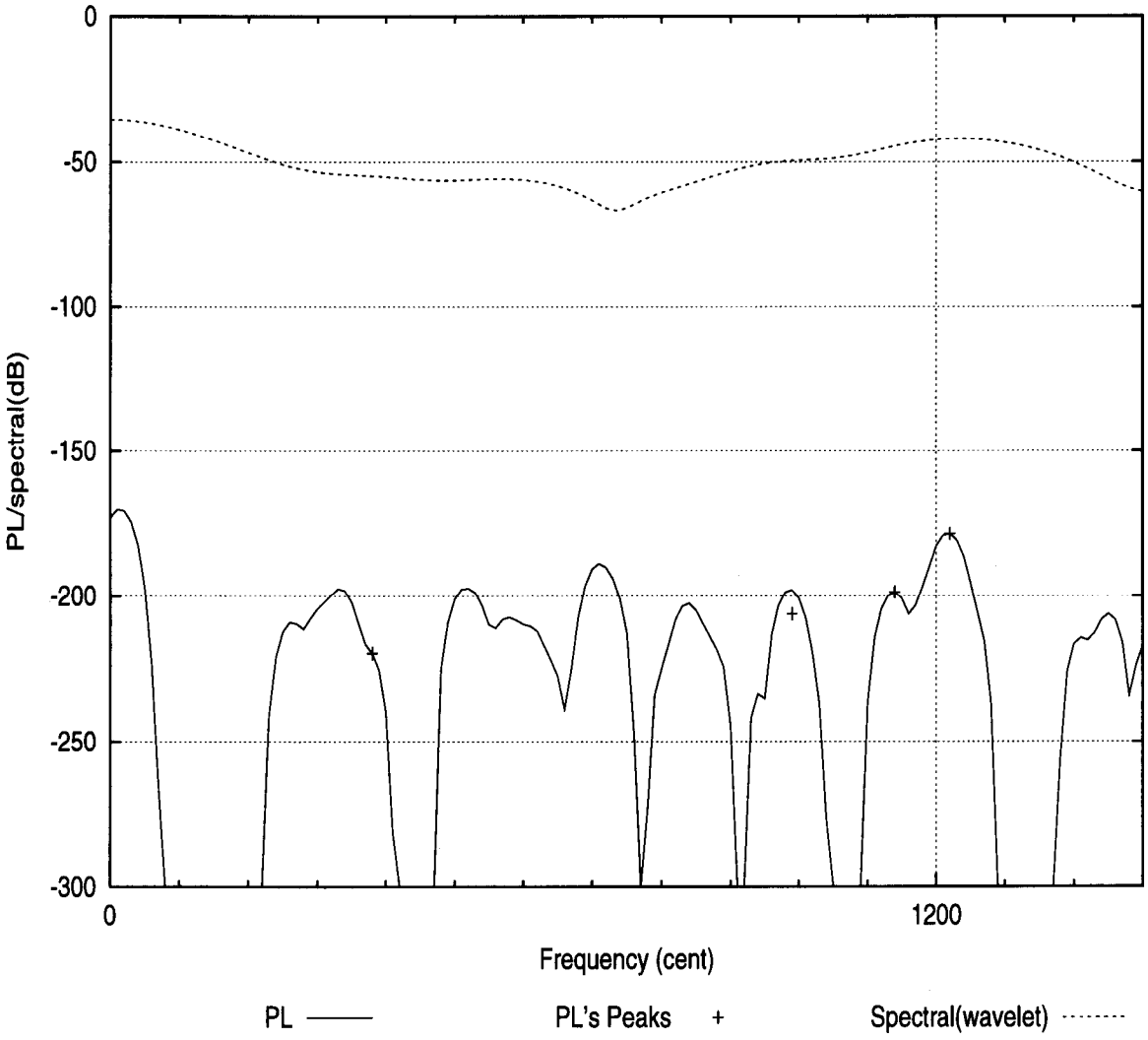

図 2.25: 図 2.24 を拡大したもの

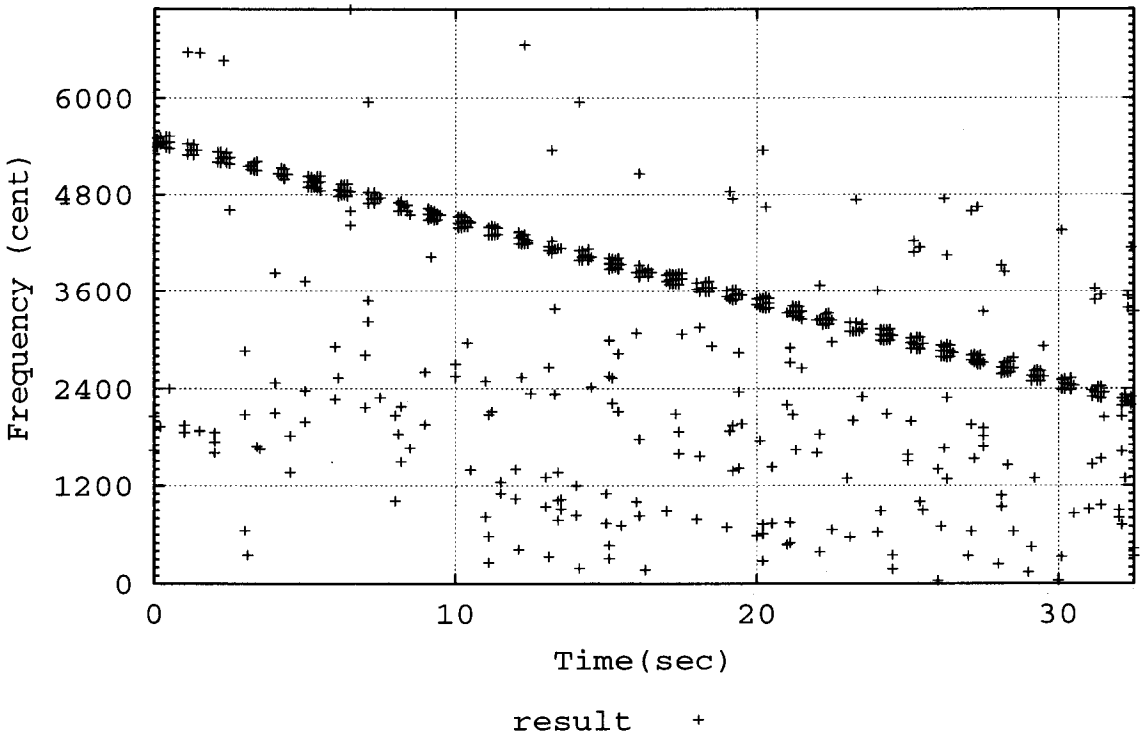

図 2.26: 複音実験 1 の結果 (可変長 DFT+Multeapics)

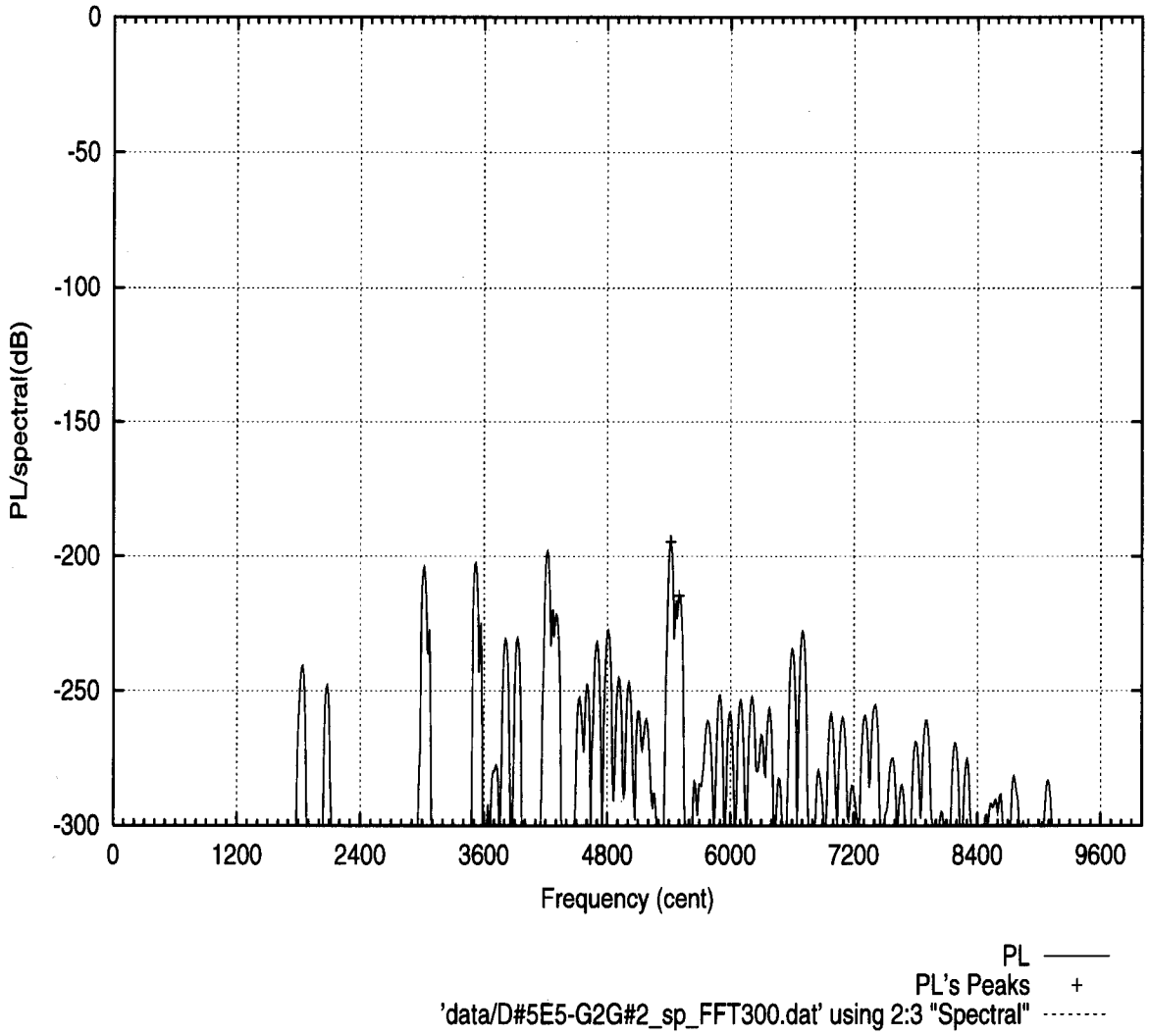

図 2.27: 複音実験 1(300msec)FFT+Multeapics

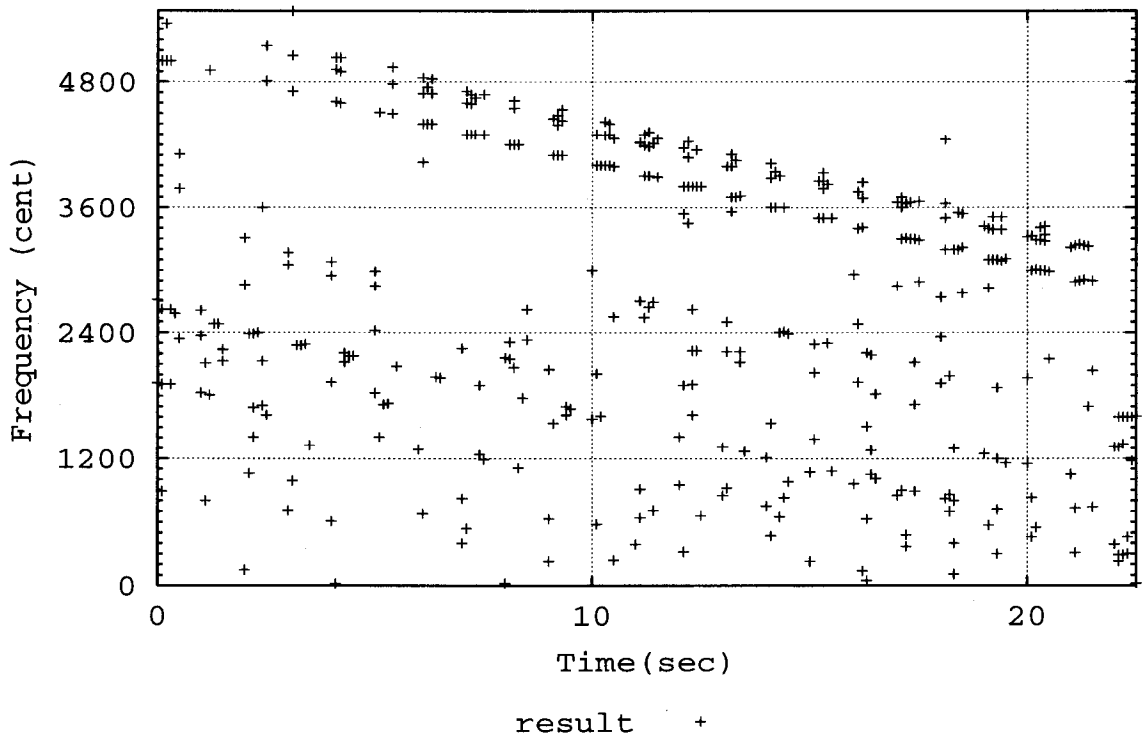

図 2.28: 複音実験 2 の結果 (FFT(可変)+Multeapics)

第3章　うなりを利用したヴォイシングの同定

3.1　近接ノートの識別

ジャズピアノでは、全音(200cent)や半音(100cent)の差で複数の音がなる近接音が多い。これらを単純にFFT等の周波数解析によって求めようとすると、不確定性により必要な分解能は得られない。そこで、時間分解能を優先して周波数解析を行いうなりを用いて近接ピークが存在するかを識別し、それらの情報を統合することで近接音を識別する手法を提案する。

3.2　理論

周波数の近い2つの波が存在し、これらが

$$x_1(t) = \sin((\omega + \Delta\omega)t)$$
$$x_2(t) = \sin((\omega - \Delta\omega)t)$$

と書けるとき、これが混ざった波形は

$$x_1(t) + x_2(t) = 2\sin(\omega t)\cos(\Delta\omega t) \tag{3.1}$$

と表せる。ただし $0 < \Delta\omega << \omega$ とする。式(3.1)は角周波数 ω の波が、$|\cos\Delta\omega t|$ で表される包絡にしたがって振幅を変動させている、と捉えることができる。本論文ではこの包絡の周期を検出することにより、$\Delta\omega$ を算出する。

3.3　具体例

図3.1にうなりが観察される例を示す。この例は、ジャズのピアノトリオの音響信号に対し1024点FFTを220点ずつかけた結果である。1binに対応する周波数帯幅は約43.0Hzとなる。白丸で示した箇所で、200msecにわたりパワーが激しく変化しているのが分かる。単純にパワーの閾値処理で音が出ているかの判別を行った場合、同じ音が何度も出ているように分析されることが分かる。

この時刻において、実際にはピアノのD4とE♭4の音が同時に鳴らされており、D4、E♭4の周波数はそれぞれ293.66Hz(4100 cent)、311.13Hz(4200 cent)である。

しかし、それぞれの周波数ピークは埋没し2つのピークが存在することは確認できない。

図3.1よりパワーの振動は258Hzから387Hzにおいて200msecの間に4周期の変動が見られる。また559Hzから688Hzにおいても200msecの間に7周期分の変動が見られる。

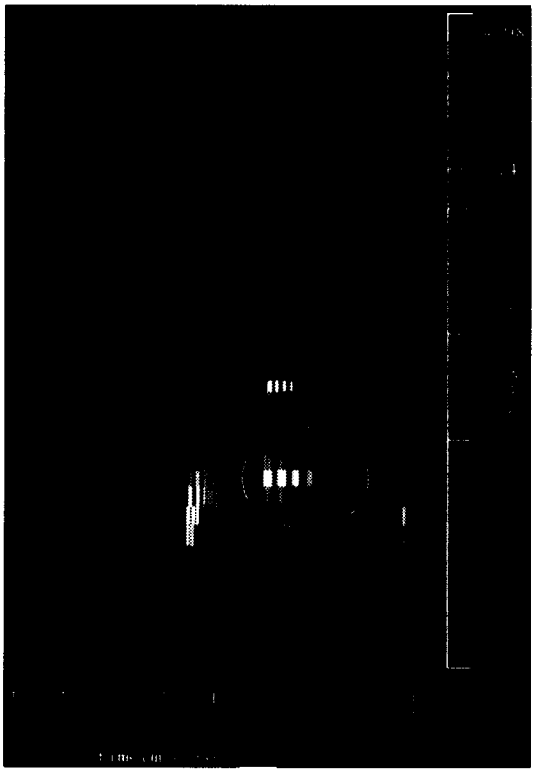

図 3.1: うなりがパワーの時間変動として現れている例(「Autumn Leaves(Portrait In Jazz)」Bill Evans)-1024 点 FFT(ハニング窓) の結果。窓のシフト幅：220 点)

4/0.2=20Hz, 7/0.2=35Hz となることから、パワーの変動から 2 組のピークの周波数差はそれぞれ 20Hz および 35Hz と推測される。

これらのノートの基本周波数差は 17.47Hz、第 2 次高調波同士では 34.94Hz と計算できるが、パワー変動の周期もほぼこの値に近い。この結果から、うなり周波数が実音響中でもパワー変動として観測されることが期待できる。

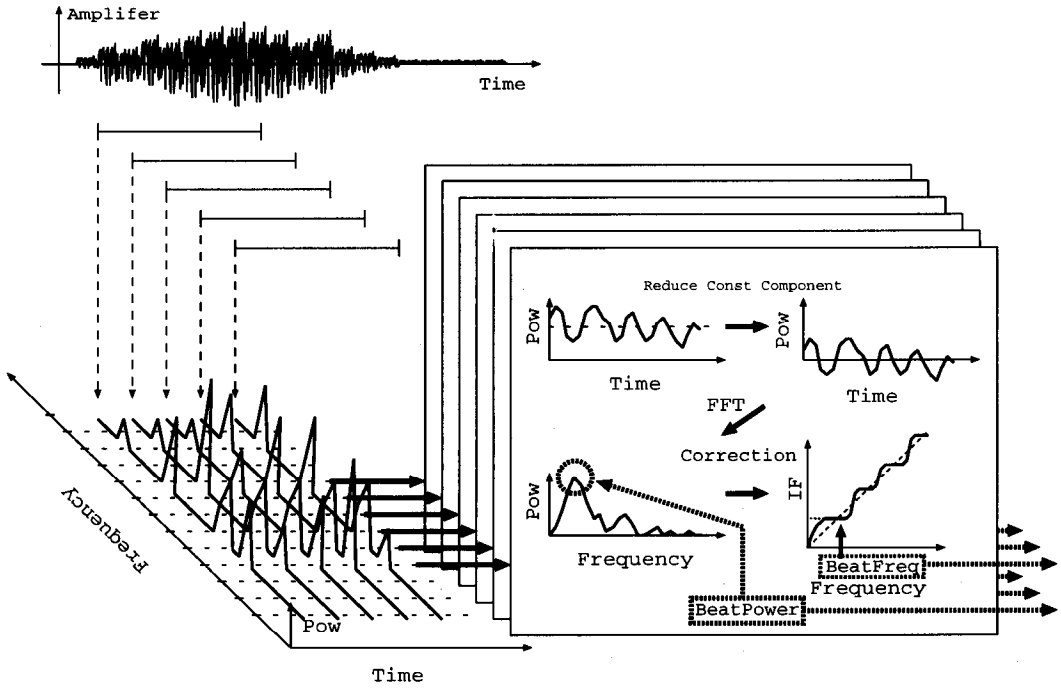

図 3.2: うなりを検出する手法

3.4 アルゴリズム

具体的な手法を図 3.2 に示す。まず、時間波形 $f(i)$ $(-1 \leq f(i) \leq 1$ で正規化) に対し STFT を用いてスペクトルを得る。

$$X(k;n) = \frac{2}{N_X} \sum_{i=0}^{N_X-1} w(i) f(i+n) W_{N_X}^{ik}$$

ここで、$W_N = \exp(-\mathrm{j}2\pi/N)$ とおいた。$w(i)$ は窓関数で、本手法ではハニング窓を使用した。$N_X = 1024$ とした。これを 1 サンプルずつシフトさせながら N_Y 回計算した後 k を固定し n を時間軸とする波形と見てもう一度 STFT を行う。本手法では、$N_Y = 4096$ とした。この際 $X(k;n)$ には定常成分が多く含まれると考え、窓の区間で平均値を求めそれを引き去ったものを波形とみて演算を行った。式で表せば、

$$Y(l,k;n) = \frac{2}{N_Y} \sum_{i=0}^{N_Y-1} w(i) u(k,i;n) W_{N_Y}^{il}$$

とかける。ここで、

$$P_{avg}(k;n) = \frac{1}{N_Y} \sum_{i=0}^{N_Y-1} |X(k,i+n)|$$
$$u(k,i;n) = |X(k;i+n)| - P_{avg}(k;n)$$

とおいた。以降 $Y(l,k;n)$ を bin スペクトルと呼ぶことにする。次に $Y(l,k;n)$ からピークの存在する帯域と、ピークの正確な周波数を求めるために $Y(l,k;n)$ には、近接ピークが存

在しないという仮定をおき、瞬時周波数 [4] を用いて補正を行う。具体的には、$Y(l,k;n)$ を1サンプルずらして測定した

$$Y'(l,k;n) = Y(l,k;n+1)$$

を用い、$Y(l,k;n)$、$Y'(l,k;n)$ の位相の変化から瞬時周波数を計算する。

$$Y(l,k;n) = a + \mathrm{j}b$$
$$Y'(l,k;n) = a' + \mathrm{j}b'$$

とおけば、位相の変化 $\Delta\theta(l,k;n)$ は内積の定義から、

$$\Delta\theta(l,k;n) = \arccos\left(\frac{aa' + bb'}{\sqrt{a^2+b^2}\sqrt{a'^2+b'^2}}\right)$$

と表せる。これを用いて瞬時周波数 $\tilde{f}(l,k;n)$ は

$$\tilde{f}(l,k;n) = \frac{f_s \Delta\theta(l,k;n)}{2\pi}$$

表せる。ここで、サンプリングレートを f_s とおいた。また、各 $Y(l,k;n)$ に対応する中心周波数は、$\frac{l}{N_Y} f_s$ で表せるので、

$$\tilde{f}(l,k;n) > \frac{l}{N_Y} f_s$$
$$\tilde{f}(l+1,k;n) < \frac{(l+1)}{N_Y} f_s$$

を満たすような $l = l_{peak}$ を求め、ピーク周波数 $f_{peak}(k;n) = \tilde{f}(l_{peak},k;n)$, パワー $P_{peak}(k;n) = |Y(l_{peak},k;n)|$ とおく。ピークのうち最も低い周波数をもつものがその帯域のうなり周波数であるとした。その後

$$P_{dB}(k;n) = 10\log_{10} P_{peak}(k;n)$$

とおき、ピークの集合 $S_{peak}(n)$ を

$$S_{peak}(n) = \{(\frac{k}{N_X} f_s, f_{peak}(k;n), P_{dB}(k;n)) \mid P_{dB}(k;n) > P_{th}\}$$

とする。本論文では $P_{th} = -50\mathrm{dB}$ とした。更に2つのピーク $(f_1, \Delta f_1, P_1), (f_2, \Delta f_2 P_2) \in S_{peak}(n)$ が

$$|f_1 - f_2| < \Delta f$$
$$P_1 > P_2 \tag{3.2}$$

を満たすときは、スペクトルの漏洩によって生じたものと判断し (f_2, P_2) を無視する。以上の処理により、ある時刻におけるピーク周波数とそのパワーを決定した。

3.5 実験1：合成音での実験

3.5.1 実験条件

合成音を使い本手法の有効性を評価する。1つのノートを

$$f_{note}(i) = \sum_{n=1}^{N} \sin(2\pi n f i / f_s)$$

としてモデル化し、$N = 8$ として合成を行った。このノートを2つ用意し足し合わせて音程を作る。ノートが低くなるに従って基本周波数は近接し、判別は難しくなる。ここでは、音楽的妥当性として Low Interval Limit[8] (コードとしてのサウンドの明瞭さを失わない最低音。以後 L.I.L と略す。) に従っているものとする。

L.I.L は音程により変わり、表3.1 のようになっている。表中の度数表記で m,M,P,+はそれぞれ、短, 長, 完全, 増に対応し、Non は最低音の制限がないことをあらわす。最高音については、高い方のノートがピアノの最高音になるように定めた。これらの条件を満すもののうち完全1度と8度を除いた音程と音域の組合せ877音について、評価を行った。

表 3.1: Low Interval Limit

音程	最低音	音程	最低音	音程	最低音
m2	E3	P5	A#1	M9	D#2
M2	C3	m6	F2	m10	C2
m3	A#2	M6	D#2	M10	A#1
M3	G#2	m7	D#2		
P4	D#2	M7	D#2		
+4	F2	m9	E2		

ノートナンバー N_1, N_2 の i_1 次高調波と i_2 高調波の周波数差は f_{A4},N_{A4} をそれぞれ A4 のピッチ、A4 のノートナンバーとすると、$f_r(n) = f_{A4} \times 2^{\frac{n-N_{A4}}{12}}$ を用いて

$$\Delta f_{\text{理論値}}(n_1, n_2, i_1, i_2) = |i_1 f_r(n_1) - i_2 f_r(n_2)|$$

とかける。これを n_1, n_2 すべての組合せについて計算し、中心周波数を $((i_1 f_r(n_1) + i_2 f_r(n_2))/2)$Hz とし、$\Delta f_{\text{理論値}} < \Delta_f$ を満たすものを理論値として用いた。本論文では $\Delta_f = 50$Hz とした。

3.5.2 評価方法

評価のために結果のクラス分けを行った。まず

1. 検出し理論値でも存在した
2. 検出しなかったが理論値では存在した (C1)
3. 検出したが理論値では存在しなかった (C2)

の3条件に分ける。これらの判別のために

1. 理論値、実験値の中心周波数の差が f_{center_delta} より小さくなるすべての組合せを作る
2. 1. について、1つの理論値に対し複数の実験値が対応している場合、中心周波数差が大きいものを削除
3. 2. について、1つの実験値に対し複数の理論値が対応している場合、中心周波数差が大きいものを削除

という処理を行った。対応の取れなかったものをC1,C2に分類する。本論文では $f_{center\_delta} = 50\text{Hz}$ とした。更に1. について

1. うなりの周波数も近かった (C3)
2. うなりの周波数は遠かった (C4)

の2つに分ける。$|\Delta f_{\text{理論値}} - \Delta f_{\text{実験値}}| < f_{delta\_delta}$ を満たすものをC4、満たさないものをC3と分類する。本論文では $f_{delta\_delta} = 1.0\text{Hz}$ を用いた。

C1～C4に属する組合せの個数を数えることで評価を行った。

表 3.2: 合成音での実験結果

	C1	C2	C3	C4
個数	530	71	96	576

3.5.3 実験結果および考察

結果を表 3.2 に示す。これは正しく認識したものが 49%と悪い。しかし、実験値が存在した場合のみを考えてみると、うなりの有無とその周波数の両方を検出した割合は 79%、うなりの有無だけであれば検出率 91%となり、本手法が近接ピーク検出に有効であることが示された。

この周波数は実験値を出すために使用した音響データ 4096 + 1024 点分のデータから解析できる最小周波数は $44100/(4096+1024) = 8.61$Hz となるが、C1 に入った理論値 530 個のうち、262 個はうなり周波数の理論値が 8.61Hz を下回っていることから、原因は不確定性であることがわかる。N_Y をより増やして測定することで、このエラーを無くすことができると思われる。

また、C2 に属する実験値の中に式 (3.2) で行ったスペクトルの漏洩の除去が失敗していることが原因であるものが含まれていた。

これは、例えば f_1,f_2,f_3 を中心周波数とする 3 つの帯域のうなりの周期のパワーが 1 組の近接ピークの漏洩によって p_1,p_2,p_3 となっているときは f_1 以外は除去されることを期待したが、$p_2 < p_3 < p_1$、そして p_2 が式 (3.2) を満たし、かつ f_1 と f_3 が式 (3.2) を満たさない場合、f_2 は閾値処理で消され、f_3 は周波数の近いピークが存在しないことから漏洩の影響ではないと判断され、除去されずに残る。このようにして、C2 へ属しているピークが幾つか存在した。

3.6 実験2：ピアノのみが鳴っている実音響での実験

表 3.3: Bessie'sBlues: ノート開始時刻, 終了時刻, 単/和音を記録 (検出元)

id	start (sample)	end (sample)	単/和音
1	47200	57600	和
2	57600	64300	単
.	.	.	.
.	.	.	.

表 3.4: Bessie'sBlues: ノートの開始時刻, 終了時刻, 単/和音を記録 (検出結果)

id	start (sample)	end (sample)	center(Hz)
1	49000	55000	650
2	93000	96500	1000
.	.	.	.
.	.	.	.

表 3.5: 表 3.3 と表 3.4 の対応をとったもの

note start (sample)	detect start (sample)	detect end (sample)	note end (sample)	単/和音
47200	49000	55000	57600	和
92400	93000	96500	104600	和
172000	175000	182000	183000	単
243000	244900	245050	250500	和
256000	257000	262000	263000	和
272000	273000	275000	290000	和
272000	273050	277500	290000	和
325000	327000	338000	339500	和
340500	343000	345000	352000	和
387000	388000	396000	399000	和
431000	432000	444100	453000	和

次に実音響において実験を行った。対象として選んだのは「Bessie's Blues」(Chick Corea Acoustic Band) の冒頭 10 秒間でピアノのイントロの一部である。これを、各時刻で周波数帯にかかわらず近接ピークが出現したかで評価を行う。

この曲を選んだ理由としては、同時に 1 音しかなっていないつまり、音は出ているが近接ピークが存在しない時刻が多く存在すること、和音はすべて不協和で弾かれており、近接ピークが存在することが挙げられる。以下に手順を示す。

1. 単音または和音が鳴っている時刻を記録する (図 3.3)
2. 実験値を求める
3. 2. を時間、周波数でプロットしグラフからクラスタの開始時刻、終了時刻を記録 (図 3.4)
4. 3. と 1. を時間関係から対応をとる (図 3.5)

1. に関しては筆者自身が採譜を行い、それに基づいて時間波形と音をたよりに記録した。

3.6.1 結果

図3.5 より、1 つを除き本手法により近接ピークを検出した時刻の 94%で採譜結果が和音であり、近接ピークを正しく検出していることが分かる。

3.7 考察

3.7.1 窓のシフト幅

本実験では、うなりが起こっているときの$Y(n)$のピークは実測値から-50dBという閾値を用いた。実際にはどのくらいの影響があるか調べる。

今

$$g_1(t) = a_1 cos(2\pi f_1 t) \tag{3.3}$$
$$g_2(t) = a_2 cos(2\pi f_2 t) \tag{3.4}$$

で表される波が重なった場合を考えてみる。それぞれの位相を考えていないのは、位相がずれている状態はこれらの波を時間的にシフトすることで得られるからである。

$a_1 = a_2 = 1.0, f_1 = 10.1, f_2 = 10.4$ の場合の波形を図3.3に示す。

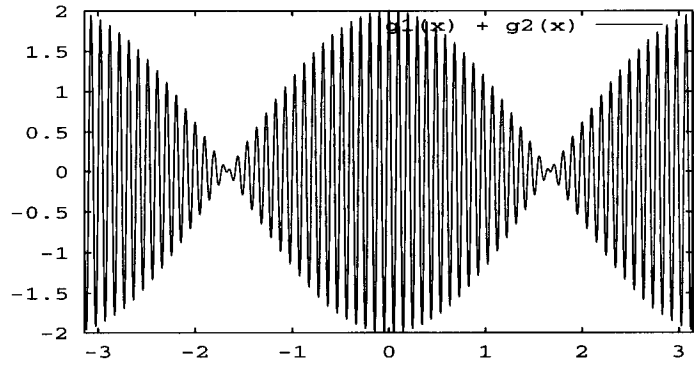

図 3.3: 式(3.3)の波が重なったもの $a_1 = a_2 = 1.0, f_1 = 10.1, f_2 = 10.4$

これより、最も大きい振幅として$a1 + a2$を、最も小さいときの振幅が$a1 - a2$であることが予想される。

周波数の大小と振幅の大小で4つの場合に分けて考える。

1. $f_1 < f_2, a_1 < a_2$(図3.4)
2. $f_1 > f_2, a_1 < a_2$(図3.5)
3. $f_1 < f_2, a_1 > a_2$(図3.6)
4. $f_1 > f_2, a_1 > a_2$(図3.7)

これらより、うなりの幅については、小さい方のパワーが崩落線のパワーとして現れると考えられる。

小さい方のパワーが崩落線のパワーとして現れるとすると、実験において、-50dBの閾値を設けるということは、波のパワーが-50dB以上のものだけを残した、ということになる。

-50dB以下の音を無視して良いかを調べる。

$$y(t) = a \exp \mathrm{j}(2\pi f \frac{t}{T} + \theta) \tag{3.5}$$

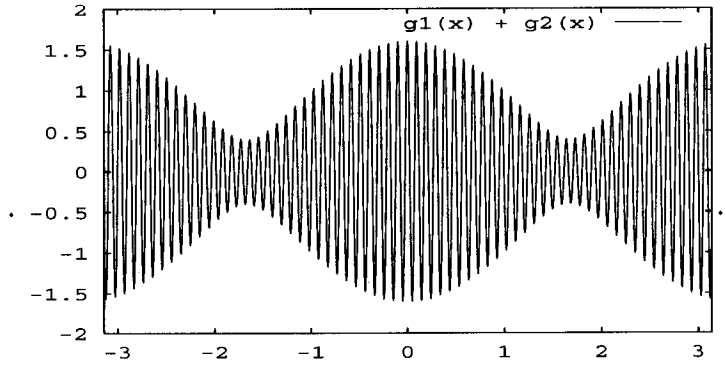

図 3.4: 式 (3.3) の波が重なったもの $f_1 < f_2, a_1 < a_2$

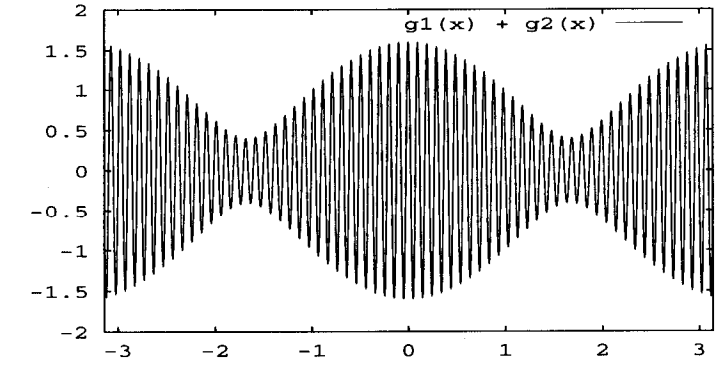

図 3.5: 式 (3.3) の波が重なったもの $f_1 > f_2, a_1 < a_2$

という波形が存在したとき、矩形窓を使った FFT による結果 $F(n)$ は

$$\begin{aligned} F(n) &= \int_{-\infty}^{\infty} w(t)y(t)\exp(-\mathrm{j}2\pi n\frac{t}{T})dt \\ &= \int_{-T/2}^{T/2} y(t)\exp(-\mathrm{j}2\pi n\frac{t}{T})dt \\ &= \frac{\sin\pi(f-n)}{\pi(f-n)}aT\exp\mathrm{j}\theta \end{aligned} \tag{3.6}$$

と表せる。これは、$F(n)$ の位相は (符号の反転による π ずれは抜かして) θ だけに依存し n には依存しないことを表す。$F(n)$ のグラフを図 3.8 に示す。ここで、$f = 30, a = T = 1$ として算出した。

ハニング窓は

$$\begin{aligned} w(t) &= \frac{1}{2}(1+\cos 2\pi\frac{t}{T}), t \in [-\frac{T}{2}, \frac{T}{2}] \\ &= \frac{1}{2}(1+\frac{1}{2}\exp(\mathrm{j}2\pi\frac{t}{T})+\frac{1}{2}\exp(-\mathrm{j}2\pi\frac{t}{T})) \end{aligned} \tag{3.7}$$

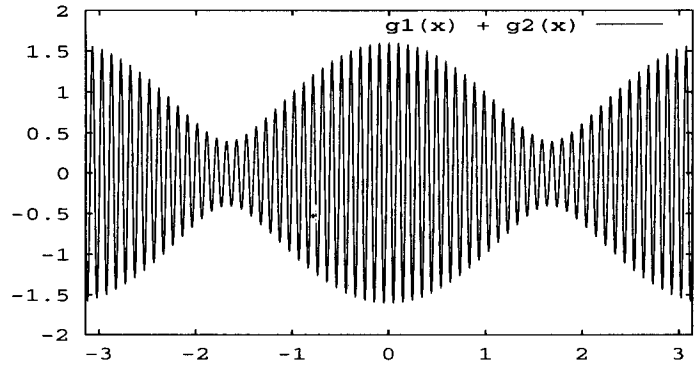

図 3.6: 式 (3.3) の波が重なったもの $f_1 < f_2, a_1 > a_2$

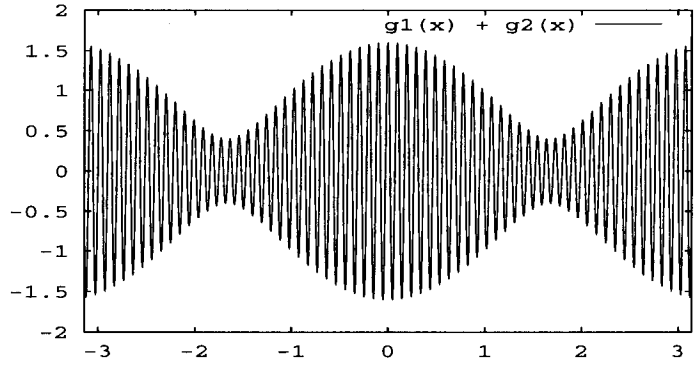

図 3.7: 式 (3.3) の波が重なったもの $f_1 > f_2, a_1 > a_2$

と表される。これより、

$$\begin{aligned} H(n) &= \int_{-\infty}^{\infty} w(t)y(t)\exp(-\mathrm{j}2\pi n\frac{t}{T})dt \\ &= \frac{1}{2}\int_{-T/2}^{T/2} y(t)\exp(-\mathrm{j}2\pi n\frac{t}{T})dt \\ &\quad +\frac{1}{4}\int_{-T/2}^{T/2} y(t)\exp(-\mathrm{j}2\pi (n+1)\frac{t}{T})dt \\ &\quad +\frac{1}{4}\int_{-T/2}^{T/2} y(t)\exp(-\mathrm{j}2\pi (n-1)\frac{t}{T})dt \\ &= \frac{1}{2}F(n)+\frac{1}{4}F(n-1)+\frac{1}{4}F(n+1) \end{aligned} \tag{3.8}$$

これを図 3.9 に表す。ただし、最も高い $f = 30$ における値を 0dB にするために、式 (3.8) を 2 倍したものを用いた。

図 3.9 の中心付近を拡大したものが図 3.10 である。この図からは

1. 実際の周波数から 2bin 分離れれば、-40dB 以下になっている
2. 実際の周波数から 4bin 分離れれば、-50dB 以下になっている

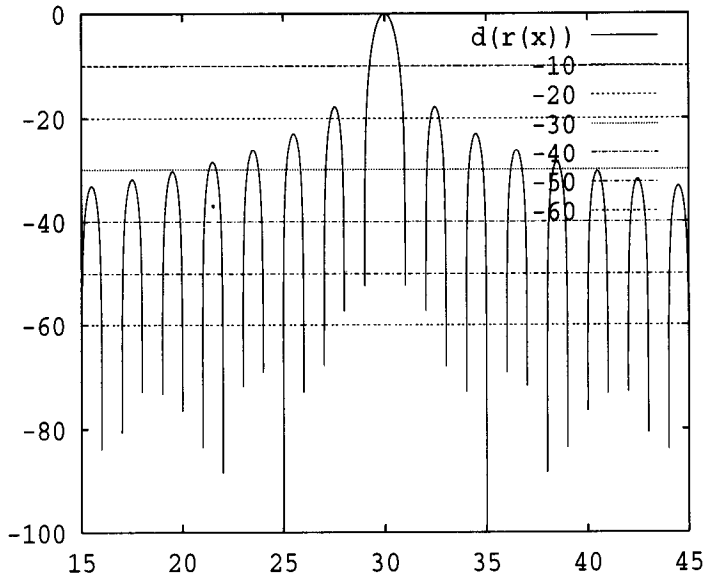

図 3.8: 方形窓によるスペクトルの漏洩

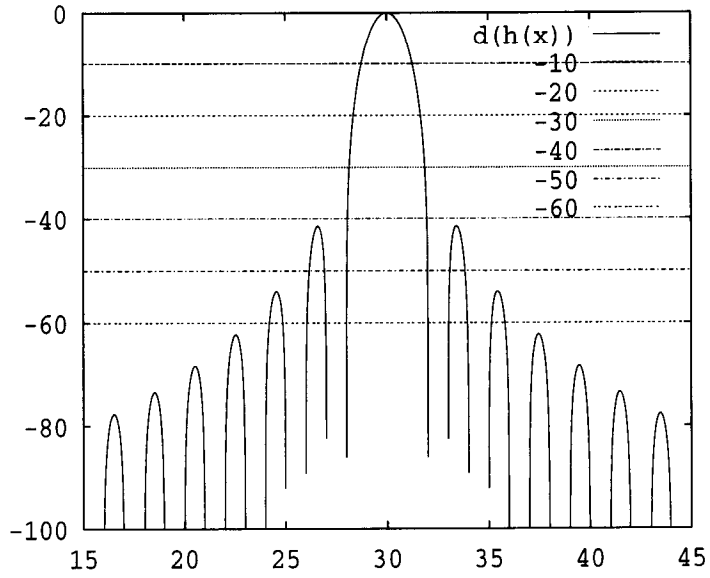

図 3.9: ハニング窓によるスペクトルの漏洩

3. 実際の周波数から 6bin 分離れれば、-60dB 以下になっている

ということが読み取れる。

設定していた閾値は-50dB なので、これは、「どんなに強いパワーでうなりが鳴っていてもその影響は 4bin 離れれば、うなりと判定することはなくなる」ということである。つまり (区間周波数) × 4Hz より大きい周波数差が検出されることはない。存在するピーク周波数の上限が決められているということは、窓のシフト幅は上限値に対応する周期の 2 分の 1 で良いことになる。

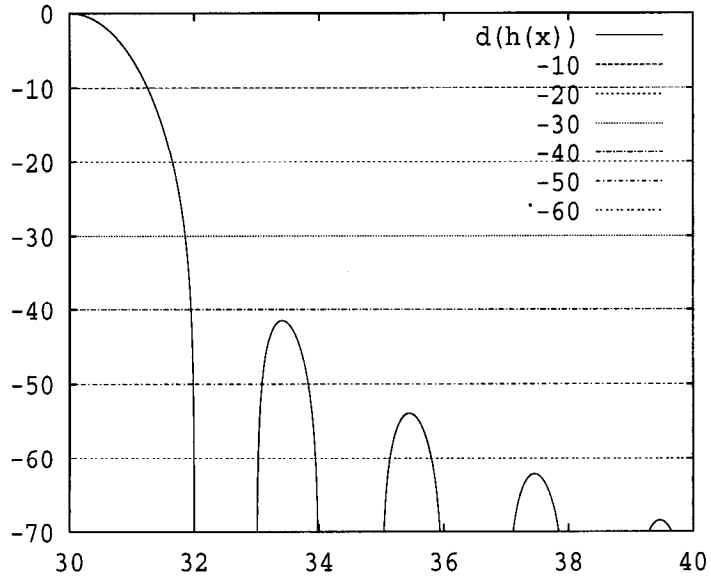

図 3.10: 方形窓によるスペクトルの漏洩

ハニング窓を使った場合、ピーク周波数から 4bin 以上離れれば必ずパワーは -50dB 以下になる。

合成音について行った実験では 0〜50Hz 内に入ったもののみを使用していた。FFT を 4096 点でかけた場合 50Hz は約 4.6bin となる。合成音ではすべてのパワーが同じだったので、周波数差の条件をきつくする為にはたらいており、しかも周波数条件はもっときつく設定してあったので、意味がなかったのではないか？と考えられる。

設定すべき問題となるのは、「成分がない」とみなすパワーは何 dB かということになる。

-50dB の折り返し歪みは無視できるものとすると、上で述べた区間と同じように 4bin 程度で良い。よってうなり判別の為には $S_r \times 4 \times 2/N_{FFT}$ Hz のサンプリング周波数以上であれば、十分である。この周期でサンプリングするためには、N_{FFT} 点 FFT のシフト幅を $N_{FFT}/8$ 点にすれば良いことが分かる。

3.7.2 システムへの実装

実際のシステムとしてボイシングを当てるためには

1. どの帯域でうなりが起きているか
2. どの周期でうなりがおきているか
3. どのくらいの振幅でうなりが起きているか

という局所的な周波数の差情報のみでは、完全にボイシングを記述することはできない。和音進行を事前に得ておくことで、ボイシングの選択肢を絞り込むことができると考えられる。

音響情報から和音情報を抽出する研究はあるものの、文献 [12] や [1] 等があるが、それらはクラシック音楽を対象にしており、本研究のような打楽器がビートを刻んでいるような対象ではない。

特に [1] のようなスペクトルパターンから直接和音情報を取り出す手法では、
本研究の対象としているジャズ音楽では和音が短い音価で複雑なタイミングで鳴っているようなものが対象では、うまく動作しないと思われる。
このことから、有用な情報を持っているのは、ベースの演奏であると考えられる。
ベース奏者が実際に演奏するノートを選ぶ際には、与えられれたコード情報に加えて

1. 今から弾くのが小節のうち何拍目の音か
2. 次の小節の和音は何か

という情報への依存が大きいと思われる。
文献 [12] では、単音から和音情報の生成を行っている。その際に既存の譜面の統計データから得られた条件付き確率と和音遷移情報を用いている面では精度向上が見込めるものの、演奏されたノートが小節中の何拍目、あるいはコード遷移の何拍前なのかを加味していない。
ジャズ音楽においてベースによって演奏されるノートは何拍目を演奏しているかに強く影響を受けている。
よって本研究の対象であるジャズに対する認識においてはテンポ、小節等の時間情報が重要な意味を持っていると思われる。
まず時間情報を確実に捉える方法を探る必要があると考えられる。

第4章　まとめ

ピッチ周波数同定法、近接音の識別法についての評価を行った。

ピッチ周波数推定法 Multeapics 法においては、単音認識、2 音の近接ノートについては高い精度が出ている。しかし、音数が増えるにつれ、精度が下がって行く傾向がある。

うなりを用いた近接音の識別では、有用な情報が得れるものの、ある 1 箇所のみの情報のみでは完全にボイシングを識別することはできない。完全に記述するためには和音情報が非常に有用と考えられる。そのためには、ベース音から和音情報を得る必要があるが、ベース音は時間情報への依存が強く、時間情報をまず求める必要がある、と考えられる。

各手法とも有用な情報を得られることを確認したが、完全にノートを識別するには至らない。

より精度の高い認識を行うためには、音響情報からの演奏が行われた後の情報のみを考えるのではなく、和音遷移等の楽典規則や演奏者の癖を考慮し、演奏が行われる前、すなわち演奏者がどのような意図を持って音を発するかについてモデル化をする必要があると思われる。

関連図書

[1] Takuya FUJISHIMA: Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music ,CCRMA

[2] 山浦敦: 和声理論に関する工学的検討-音程知覚過程のモデル化とその応用- , 東京大学工学部卒業論文,1991

[3] 西 一樹, 阿部 素嗣, 安藤 繁：Wavelet 空間における特定倍音成分の分離抽出, 電気情報通信学会, 信学技報 EA95-39,1995

[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, 電子情報通信学会論文誌,Vol.J79-D2,No.11,pp.1771-1781(1996)

[5] 高澤 嘉光: 離散フーリエ変換における補間公式, 音楽音響研資,MA89-26,1998

[6] 青野 勇司, 片寄 晴弘, 井口 征士: アコースティックピアノを用いたセッションシステムの開発, 情報処理学会研究報告 MUS-21,No.6,pp.31-35(1997)

[7] 片寄 晴弘著: 自動採譜 (概論) , 共立出版 bit 別冊「コンピュータと音楽の世界-基礎からフロンティアまで-」,pp.74-88(1998)

[8] 飯田 敏彦著: やさしく学べるジャズハーモニー 2 , 全音楽譜出版社,1984

[9] 吉川 茂著: ピアノの音色はタッチで変わるか, 日経サイエンス社,1997

[10] 後藤 真孝: 音楽音響信号を対象にしたメロディーとベースの音高推定, 電子情報処理学会論文誌 D-II,Vol.J84-D-II,No.1,pp.12-22(2001)

[11] 後藤 真孝: リアルタイム音楽情景記述システム:全体構想と音高推定の拡張, 情報処理学会研究報告 MUS-27,No.2,pp.9-16(2001)

[12] 柏野邦夫著: 音楽音響信号を対象とする聴覚的情景分析に関する研究, 博士論文, 東京大学工学部,1994