

## うなりを利用した不協和音程の検出

杉浦 勇樹

阪口 豊

電気通信大学大学院・情報システム学研究科

概要：本研究の目的はジャズのピアノトリオを対象とした自動採譜システムの構築である。ジャズ音楽においては不協和音が頻出し、これによりスペクトル上で近接ピークが存在する。本報告では、うなりを使って近接ピークを検出する手法を提案する。8倍音までの高調波と同じ強度で含む合成音を用いた実験では、近接ピークが存在する周波数帯を91%で、ピーク周波数の差を79%の割合で検出した。実音響を用いた実験では検出した16箇所の時刻のうち94%で近接ピークが存在した。

## Detecting Dissonance Utilizing Beat of Spectrum

Yuuki Sugiura

Yutaka Sakaguchi

The University of Electro-Communications

Abstract: Our goal is to develop a music transcription system whose main target is piano trio in jazz music. We proposed a new method for detecting the musical dissonance, which frequently appeared in jazz music, utilizing the beats. The results of two preliminary experiments, one dealing with artificial sounds, and the other dealing with sounds from piano solo, showed that the proposed method detected closely located frequency peaks and the difference between the frequencies with good scores.

### 1 はじめに

本研究の目的はジャズのピアノトリオ(編成はピアノ、ベース、ドラム)のデジタル化されたオーディオ信号からピアノ音の採譜をおこなうことである。ジャズ音楽ではアドリブ演奏が多く、演奏の数だけ譜面を作ることができるため、これを自動生成するシステムは需要が高い。

ジャズ音楽を対象にした場合、特異的な問題として、

1. 速い音の動き
2. 緊張感の高い和音

が挙げられる。本報告では後者への取り組みを行う。

一般に、周波数解析において時間精度と周波数精度にはトレードオフの関係があり、単純な解析だけでは出現するスペクトルのピークの周波数と時刻とを検出するため必要な精度は得られない。

そこで、いくつか周波数精度を補間する手法が提案されている[1][2]が、これらは近接したピークは存在しないという仮定をおくかあるいは実装上この仮定をおくことになる。

しかし、本研究で対象としているジャズのピアノトリオでは、特に、200~400Hzの中音域において近接したピークが多く、低次倍音において先の仮定が成り立たない時が多い。

そこで、1つのbin(DFTにおいて1点で表す解析区間の単位)に2つのピークが入り込んでしまった場合にこれを判定することを考える。

具体的には、周波数解析によって求めたパワーを時間的に観察しその周期を計測することで、2つのピークを検出することを考えた。尚、実験に出てくる音響信号は特に記述のない限りサンプリングレート44.1kHz、量子化ビット数16のデータを使っているものとする。

## 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

$$x_1(t) + x_2(t) = 2 \sin(\omega t) \cos(\Delta\omega t) \quad (1)$$

と表せる。ただし  $0 < \Delta\omega \ll \omega$  とする。式(1)は角周波数  $\omega$  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。本報告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。

### 2.1 具体例

図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動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<sub>b</sub>4の音が同時に鳴らされており、D4、E<sub>b</sub>4の周波数はそれぞれ 293.66Hz(4100 cent)、311.13Hz(4200 cent)である。

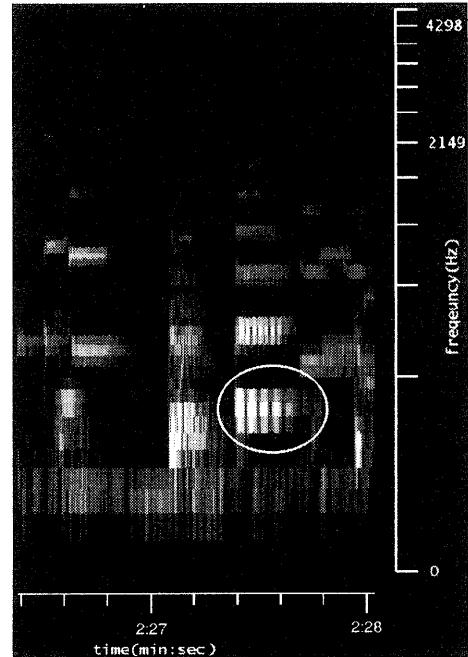


図1：うなりがパワーの時間変動として現れている例（「Autumn Leaves(Portrait In Jazz)」Bill Evans)-1024点FFT(ハニング窓)の結果。窓のシフト幅：220点)

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。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。

### 2.2 アルゴリズム

具体的な手法を図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

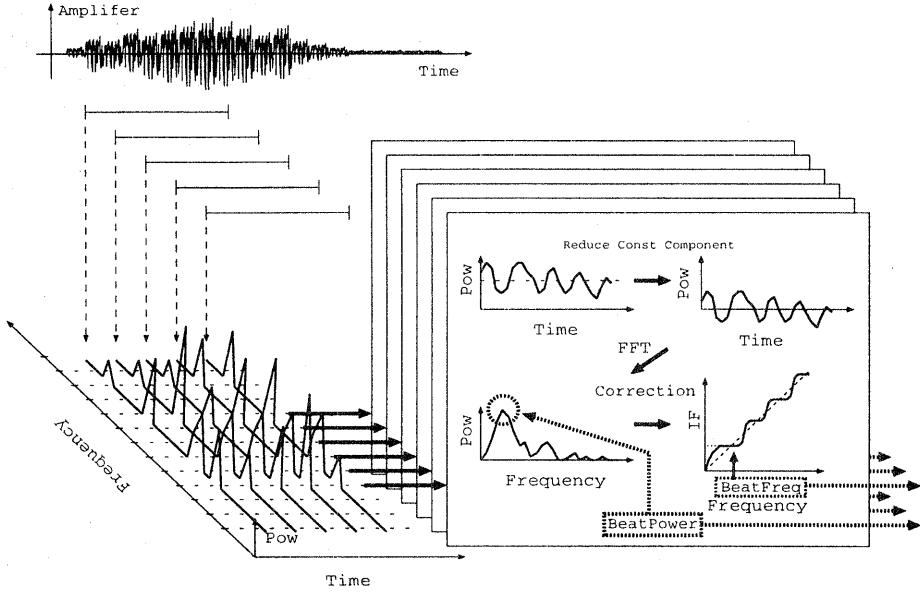


図 2: うなりを検出する手法

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、 $W_N = \exp(-j2\pi/N)$  とおいた。 $w(i)$  は窓関数で、本報告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。本報告では、 $N_Y = 4096$  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て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 [1] 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。

$$Y(l, k; n) = a + jb$$

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

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。本報告では  $P_{th} = -50\text{dB}$  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

$$|f_1 - f_2| < \Delta f$$

$$P_1 > P_2 \quad (2)$$

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 実験 1：合成音での実験

#### 3.1 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[3] (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 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。

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き、 $f_r(n) = f_{A4} \times 2^{\frac{n-N_{A4}}{12}}$  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  $((i_1 f_r(n_1) + i_2 f_r(n_2))/2)\text{Hz}$  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。本報告では  $\Delta_f = 50\text{Hz}$  とした。

#### 3.2 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

表 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		

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。本報告では  $f_{center\_delta} = 50\text{Hz}$ 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。本報告では  $f_{delta\_delta} = 1.0\text{Hz}$  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。

### 3.3 実験結果および考察

結果を表 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  $44100/(4096 + 1024) = 8.61\text{Hz}$  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  $8.61\text{Hz}$  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(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。

表 2: 合成音での実験結果

	C1	C2	C3	C4
個数	530	71	96	576

これは、例えば  $f_1, f_2, f_3$  を中心周波数とする 3 つの帯域のうちの周期のパワーが 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(2)を満たし、かつ  $f_1$  と  $f_3$  が式(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。

表 3: Bessie's Blues: 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	单
.	.	.	.
.	.	.	.

表 4: Bessie's Blues: 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
.	.	.	.
.	.	.	.

## 4 実験 2: ピアノのみが鳴っている実音響での実験

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ストロの一部である。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)
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4)
4. 3. と 1. を時間関係から対応をとる(図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。

#### 4.1 結果

図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。

#### 5 考察

本報告では、binスペクトルの計算の為に1024点FFTを1サンプルずつずらして、4096回計算した。式(1)で  $\omega >> \Delta\omega$  であるので、少なくともうなりが  $\omega$  以上の角周波数を持つことはない。近接ピークが 200Hz~400Hz に多く出現することから考えると、シフト幅を長くとり、計算量を減らすことができると考えられる。

表 5: 表3と表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

今後は、実音響において近接ピークが起こっている時刻だけでなく、周波数帯域の判別および周波数差の精度の評価や、複数の楽器が含まれる実音響を対象にして実験を行う予定である。

#### 6まとめ

うなりを手がかりにして近接ピークの有無、および周波数差を検出する方法を提案した。この手法を8つの倍音を同じ強度で含む合成音2音で音程を作り評価を行った。それに加え実音響を用いて近接ピークが存在する時刻の検出についても評価を行った。

いずれも本手法において検出したものは、良い成績が出ているがその逆についてはエラーが多い。採譜システムを構築する際にはこの手法のみを使うのではなく、有用な情報を与える1つのモジュールとして用いる予定である。

また、本手法の利点は一度短い窓で計算してしまった周波数解析データもそれを集めて再利用し、近接ピークを知ることができると捉えることもできる。

問題点として、本手法では1つの帯域に含まれる近接ピークが2つであるという仮定の下での議論であり、この仮定の有効性については検討する必要があることが挙げられる。

今後は、近接ピークの存在とその周波数差のみをだけでなく、そのベースとなっている周波数を求め、2つの波それぞれの周波数やパワーを同定するための情報を獲得できないか検討していきたい。

#### 参考文献

- [1] 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)
- [2] 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
- [3] 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