第９回　「科学の芽」賞　応募論文

人間による音声の知覚と分解

- それに表れる計算機との相違 -

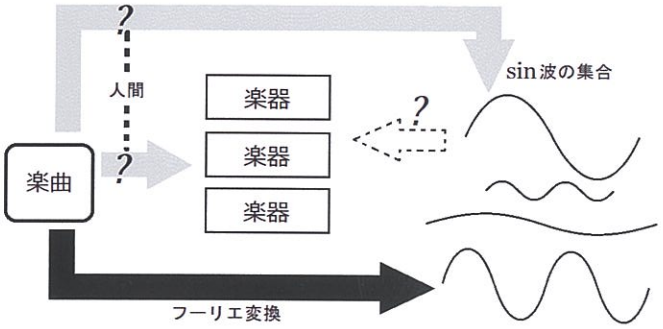
武蔵高校３年　中西 貴大

序章

我々人間は、多くの楽器の音が足しあわされた音楽を聴き、それぞれの楽器を聞き分けることができる。加えて、その分解の結果は聴くたびに一定であり、大きな個人差もないといえる。私は、正確に楽器単位に音を分解できるこの能力（システム）の拠りどころを探求しようと思った。

音は媒質の密度変化が伝播するという形態の波動であり、空間のある点における媒質の密度は時間の関数で表される。音声ファイルのひとつである WAVE において音声は、アナログ信号である音を、ある時間間隔ごと（44.1 kHz など）に量子化（サンプリング）したデジタル信号をそのまま記録するという手法で再現される。つまりそのファイルにおいては、電流 I ($\propto$変位 x) が時間 t の離散的な一変数関数 $I(t)$ で表されている。実に単純な方法だが、離散的であることを除けば、これは純粋なアナログ信号である音の形態を全く同じく表現している。

単なる一変数関数に帰着される音楽を、楽器単位に——たとえそれらが同じ音程を出していても——分解し、加えてその楽器が和音を奏でているときにはその構成音に分解するという能力が人間にはある。音の分解という解析は機械においてフーリエ変換による sin 波への分解という形で実現され、その結果はスペクトログラムとして表現されるが、[図 1]に示したようにそれは人間の分解方法とは明らかに異なる。人間の、ある意味で中途半端ともいえる音声の分解がどのようなメカニズムで行われるのかを疑問に思った。



[図 1]　音の分解と再構成

機械によって楽器単位に音声を分解しようとする技術はすでにある。時刻、周波数、そしてその強度の 3 軸から成るスペクトログラムを画像と同様に見なし（時刻を横軸、周波数を縦軸にとり、強度はふつう色調で表現されるからである）、周波数軸上の頻出パターンをいくつか抽出する、非負値行列因子分解（Nonnegative Matrix Factorization）である。抽出するパターンを基底といい、その数は楽器の数に対応する。スペクトログラムを、楽器のパターンの情報を保持した行列と、それぞれの楽器の音量の時間変化を表す行列との積に近似して分解するのが NMF である。つまり sin 波を適切に再構成する仕組みである。人間の頭脳がこのアルゴリズムを実行するわけでは決してないから、人間の頭脳の処理系、または、音を周波数成分に聞き分ける蝸牛といった人間の器官が特殊な作業を行うか、何らかの物理現象を利用しているに違いない。もし、後者の何らかの物理現象が音を楽器単位に分解するのなら、演算能力を持たない物体によって人間の分解能力が裏付けられるはずだが、特定の楽器のみに反応する物体など存在しないだろう。したがって人間の知覚と思考こそが楽器の聞き分けの方法であると考えられる。一方で実験 2 のように、音声の分解以外で人間と共通点のある挙動を示す機械あるいは環境も探すことにした。

目的

機械による数学的解析ではなく、人間による、楽器音の分解といった音の知覚の特徴を探る。

実験

事前実験　スピーカーから発せられた音声を実験に使用することの妥当性を評価する

実験 1　重ねられた等周波数の音源を人間が分解できるための条件を調べる

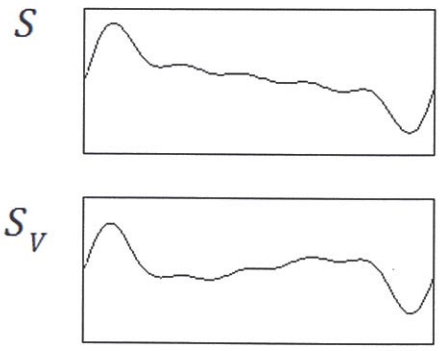
実験 1-1　　周波数と振幅が時間変化しない二つの音源を重ねた音源の分解

実験 1-2　　一方の音源の、振幅のみが時間変化する二つの音源を重ねて作った音源の分解

実験 1-3　　両方の音源の、振幅のみが時間変化する二つの音源を重ねて作った音源の分解

実験 2　バーチャルピッチに対する人間と機械の反応の差異を調べる

フーリエ変換により人間が聞く音の音色を評価することができる。たとえばピアノよりオルガンのほうが高く明るい音に聞こえ、解析結果を見るとたしかにオルガンのほうが高周波の倍音を多く含むというように、一見その結果は人間の感覚によく一致する。



[図 2]　バーチャルピッチの波形

しかし、$a_1 : a_2 : a_3 : a_4 : a_5 : \ldots$ を作為的に設定し、その音源 S と、基本振動だけを含まない（$a_1 = 0$ とした）音源 S_V を聞き比べると、S_V に存在しないはずの f_1 の周波数が聞こえるという現象がある。もちろん S_V のフーリエ変換では $a_1 = 0$ となる。この存在しないが聞こえる基本振動はバーチャルピッチとよばれる。言うまでもなくこのような倍音だけで構成された音声は自然界に存在しない。本研究では主に

$$a_1 : a_2 : a_3 : a_4 : a_5 = 10 : 8 : 7 : 5 : 3$$

なる音源 S を用いた。S と S_V の波形（[図 2]）の凹凸は大きく変わらないのがわかる。

実験 2-1　バーチャルピッチがスピーカーを原因とするものではないことの証明

我々が聞くことのできるのは作成したデータそのものではなくスピーカーから発せられる空気振動である。バーチャルピッチの信号をスピーカーに流すことで、スピーカーの何らかの性質により f_1 の周波数が生まれ、それが聞こえるのではないかという疑念を解消した。

実験 2-2　バーチャルピッチに対する音叉の反応を調べる

人間以外でバーチャルピッチに反応する実物があれば大変興味深い。共鳴実験などで用いられる共鳴箱つき音叉のバーチャルピッチに対する応答を調べた。

実験 3　楽器音の特徴を調べて応用する

実験 3-1　楽器の音を特徴づけるパラメータを探る

ある時点での周波数成分の振幅を求めてその時間変化を追うプログラムにより、楽器音を特徴づけるものが何であるかを調べた。

実験 3-2　簡易な音声合成を試みる

実験 3-1 の結果に基づいた楽器音の再現を試みた。

実験方法

D/A 変換なくデジタル信号をそのままスピーカーに流すと仮定しても、スピーカーコーンの時間変位は離散的にはなりえないから、十分に大きなサンプリングレートをもつデジタル音声は最終的に離散的でないとさえいえる。このことから、実験にデジタル音源を使用することは妥当であると判断した。実験の手段のひとつとして、音源を処理するためのソフトウェア "tWave" を作成した。ソースコードは 60kB 程度であるので、ここには掲載しない。sin 波を重ね合わせて一波長ぶんの音源を生成するのが主な用途であるが、手動入力での各サンプルの値の変更、WAVE ファイルの読み込み／書き出し、振幅方向／時間方向の拡大縮小、二音源間の加減算と乗算、そして離散フーリエ解析（Discrete Fourier Transform）を行うことができる。DFT については、もとの複素関数の形(1)を変形して実数のみとした(2)を実装した。

$$X(k) = \sum_{n=0}^{N-1} x(n) \exp\left(-\frac{j2\pi kn}{N}\right) \qquad \ldots\ldots(1)$$

複素関数 $X(k)$ は振幅($|X(k)|$)と位相($\arg(X(k))$)の情報を持つ。

j は虚数単位である。

整数 k は離散的に周波数を表すパラメータで、$0 \le k \le N$ である。

$$a(k) = |X(k)| = \sqrt{\left[\sum_{n=0}^{N-1} x(n) \cos\left(-\frac{2\pi kn}{N}\right)\right]^2 + \left[\sum_{n=0}^{N-1} x(n) \sin\left(-\frac{2\pi kn}{N}\right)\right]^2} \qquad \ldots\ldots(2)$$

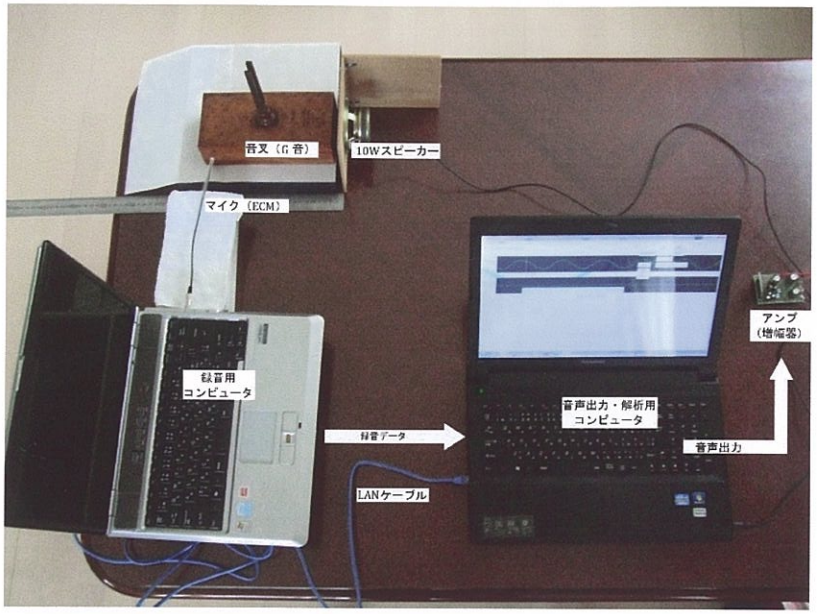
$a(k)$ は振幅である。人間の聴覚は位相の違いにかなり鈍感であることが知られており、フーリエ変換を行うソフトウェアはふつう振幅のみを結果として出力する。それにならって tWave 内に含まれる DFT を行う関数も振幅のみを出力するように作成した。DFT で解析される周波数は離散的であるが、今回の研究の一部は、周波数成分の正確なピークを求めることを必要とする。よって一般的なソフトウェアの機能に加え、手動で設定した周波数の強度の解析を行えるような関数も別途に作成した。

音声を録音しないときにはヘッドフォンを装着し、あらかじめ倍音の周波数を単独で聞いて覚えておき、それらの周波数が調べる音源の中に聞こえるかどうかを注意深く調べた。DFT の解析結果と人間の感覚が異なるような音声を発見した場合は、比較する周波数の sin 波を聞いてから、視覚においては残像にあたる弊害である耳鳴りが残っていないことを確認してから調査対象の音声を聞いた。

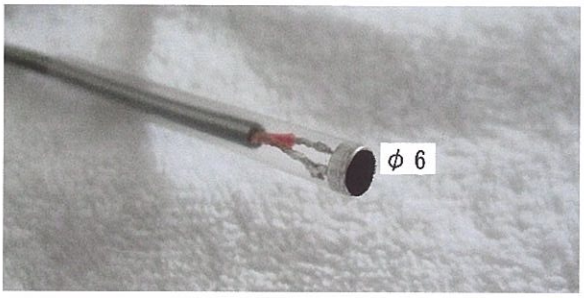
なお、以下では n 倍振動の周波数 $f_n\,(=n f_1)$、その振幅 a_n とする。時間の長さの単位としてミリ秒〔msec〕をしばしば用いた。

事前実験

スピーカーから出ている音声の波形と周波数成分とが作成したデータと明らかに異なっている場合、実験の妥当性が損なわれる。そこで、音声出力・解析用コンピュータで f_1 から f_5 までを合成した音声をスピーカーから再生し、スピーカーを取り付けてある 5mm 厚の板の表面から 10 mm のところに設置したエレクトレットコンデンサーマイク（ECM）を接続した録音用コンピュータでこれを録音し、その録音した音源を音声出力・解析用コンピュータから再生して再び録音する、という作業を繰り返し、スピーカーとマイクを通した信号の変化を追った。



[図 3]　機材の配置（実験 2-2 でのみ音叉を使用）



[図 4]　無ブランドの ECM に同軸ケーブルと 3.5 mm ミニプラグとを接続し、内径 6 mm のストローを被せたもの

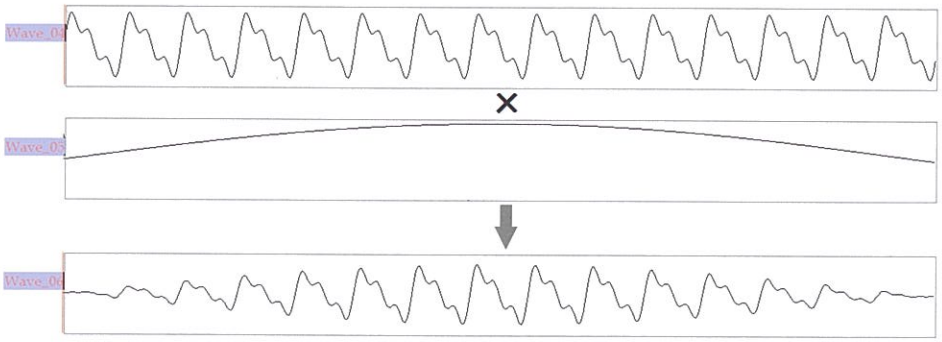
実験 1

実験 1-1

同じ周波数だが違った音色のふたつの音源 S, S′を作成する。音色が違うというのは、S, S′ の成分の振幅である $a_1 : a_2 : a_3 : a_4 : a_5 : \ldots$ の比と $a'_1 : a'_2 : a'_3 : a'_4 : a'_5 : \ldots$ の比を違うものにするということだ。S と S′ を足し合わせて作られた一波長ぶんの音源 W を繰り返して再生し、W は S と S′ とが足されたものであるとわかるかどうか調べる。

$a_1 : a_2 : a_3 : a_4 = 10 : 6 : 3 : 2$，　$a'_1 : a'_2 : a'_3 : a'_4 = 6 : 10 : 7 : 3$ なる S, S′ を用いた。

実験 1-2, 1-3



[図 5]　乗算による振幅変化イメージ

（上図でのみ視覚的に変化が明らかになるよう振幅変化の周期を 12 msec と短いものにした）

実験 1-1 で用いたのと同一の S, S′を使用した。1-2 では S′ のみの、1-3 では S と S′ 両方の振幅を、（人間が音量変化を感じる程度の周期で）時間変化させる。400 msec の周期をもつ sin 波（2.5 Hz であるため可聴域ではない）を半分の長さの 200 msec に切ったものと S とを、tWave の波形乗算機能により掛け合わせることで、振幅を sin 波に沿って 200 msec の周期で変化させるという方法で時間変化をつけた（[図 5]）。

1-3 では、S と S′の音量変化の山と谷をちょうど埋め合わせるようにし、足し合わせたときの全体としての音量変化を低減した。つまり S には $|\sin \omega t|$ を乗算するが、S′には $1-|\sin \omega t|$ を乗算するというようにである。S と S′ を足し合わせた音源 W を使って、1-1 と同様に実験した。

実験 2

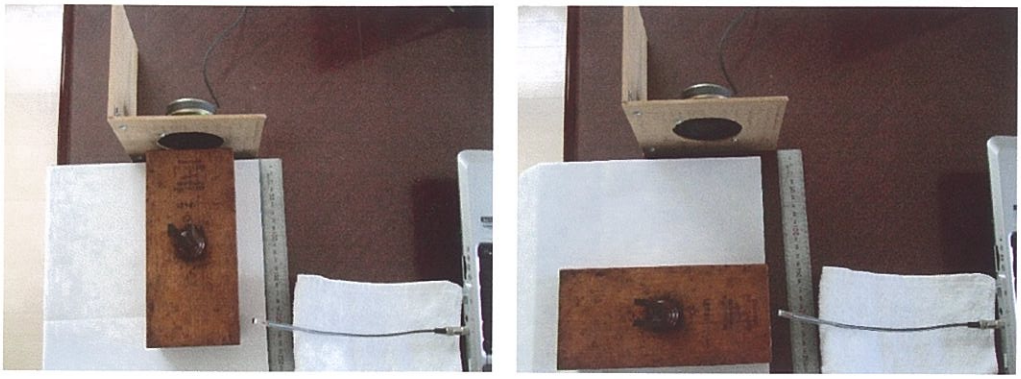
実験 2-1

比較用音源 S とバーチャルピッチ S_V をそれぞれスピーカーから再生して、事前実験のときと同じ配置のマイクで録音したものを解析した。

実験 2-2

G 音の音叉の周波数は共鳴箱に 384 Hz と表記されていたが、表面の酸化物により周波数が低下していた。基準となる電子音とのうなりを観測することで音叉の周波数 f_1=383 Hz と求めた。

[図 3]のように機材を配置し、データをとる直前に音叉に軽く触れ振動を消した。そしてスピーカーの前に設置した音叉の共鳴箱に向けて、f_1, f_2, f_3 の sin 波、比較用音源 S、バーチャルピッチ S_V という以上 5 種類を 10 sec にわたり再生したのち、音叉の共鳴があるか否かを、聴覚を頼りにせず[図 6]のように音叉の台ごと素早く静かに回転させてマイクに向け録音し、これを解析した。



[図 6]　スピーカーを取り付けた板から 10 mm の位置に開口部があるように設置した音叉

実験 3

実験 3-1

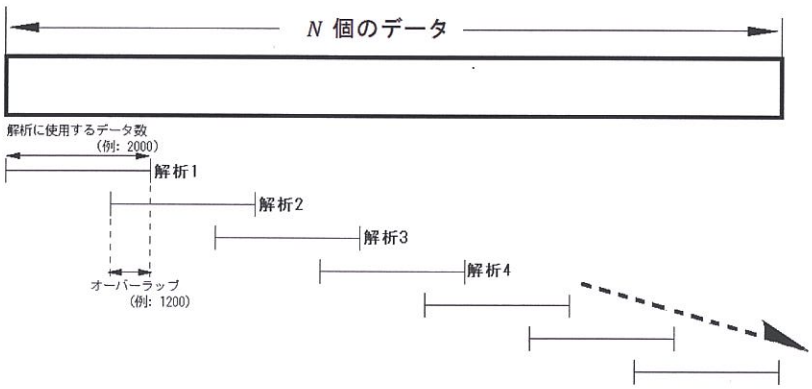
実験 1-3 から、楽器らしさを生むものは音色、つまり成分の強さの時間変化もしくは周波数そのものの時間変化であると考えられた。

楽器音に対して独自の解析を行った。一般に DFT は解析できる周波数が離散的であり、その周囲の周波数は解析可能な近い周波数に寄って算出される。解決のため、任意の周波数 f について解析できるよう(1)を(3)のように変形した（もちろん解析時には周波数は 32 ビットのデジタルデータであり離散的ではあるが、周波数の値が小数点以下 1 桁まで表現できれば十分なのである)。

$$X_0(f) = \sum_{n=0}^{N-1} x(n) \exp\left(-\frac{j2\pi fn}{f_s}\right) \qquad (f_s\text{はサンプリング周波数}) \qquad \cdots\cdots(3)$$

これを用いて、指定した周波数の振幅変化を時間軸上で走査するプログラムによって、成分の強さの時間変化を調べた。さらに複数の周波数による結果を重ねることで、もっとも強い周波数が時間とともにどう変わるかも知ることができた。

まずギターの A 音（基本振動およそ 110Hz）をマイクで録音した。次に行った DFT の結果によると、この 110Hz の音の中では二倍振動の 221Hz 周辺が最も強かったため、219.000Hz から 223.500Hz までを 0.5Hz 刻みで解析した 10 組のデータを重ね、グラフを観察した（なおグラフ描画は tWave とは別の tWaveGrapher なるプログラムに解析結果を委託することで行った)。



[図 7]　解析の概念図

解析区間をずらしながら、指定した周波数の振幅を計算してゆく。DFT の結果は解析に使用するデータ数によって少し変化する特徴があるので、この数は適切に選ぶ必要がある。

参考として、サンプリングレートが 44100 Hz のとき

- 2000 個のデータは約 0.045 sec の音声に相当する。
- 左図の設定では、10 sec の音声に対して 549 回の解析を行う。

実験 3-2

実験 3-1 からわかった楽器音の特徴を適用し、もっとも簡易な音声合成を行った。周波数特性と振幅の時間変化のみでピアノの再現を試みた。$a_1 : a_2 : a_3 : a_4 = 10 : 6 : 4 : 2$ とし、これに(4)における指数関数 V_e を掛けることで振幅を変化させた。パラメータ τ は

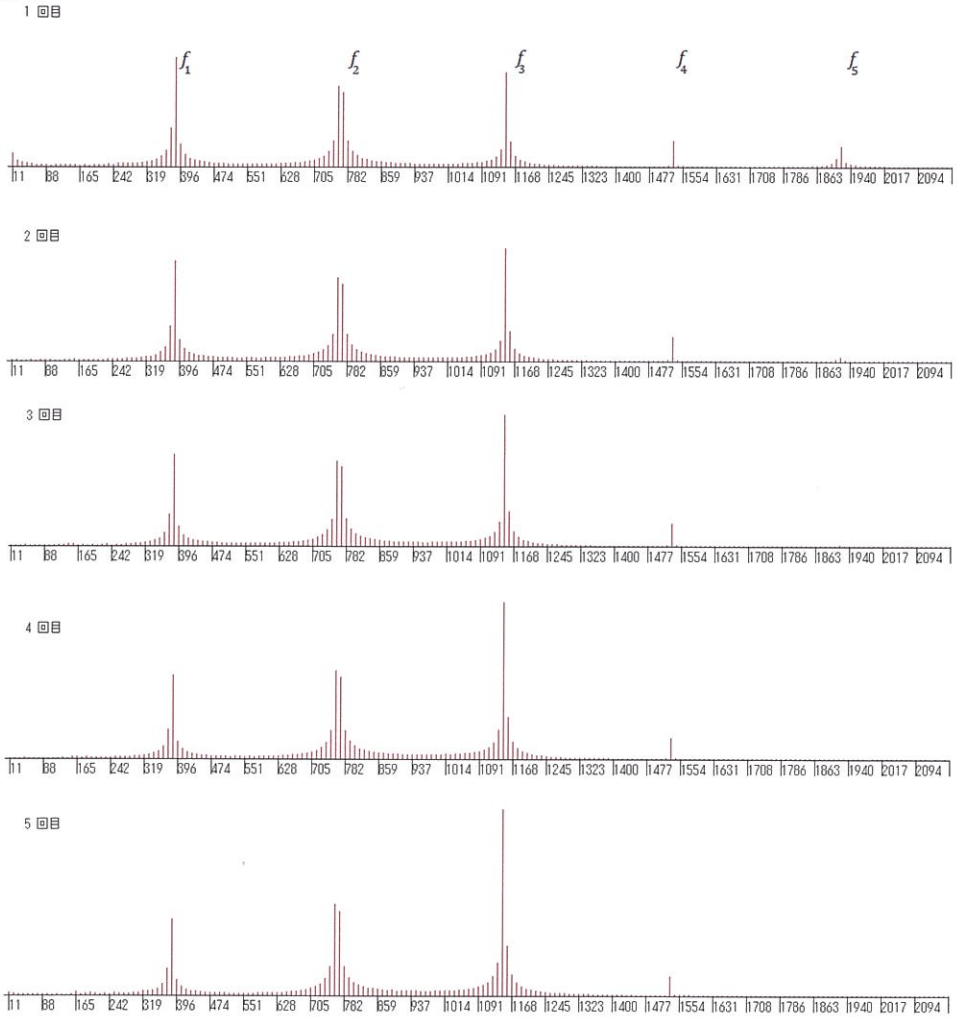
振幅が 1/e になるのに要する時間〔sec〕であり、放射性物質の半減期に似ている。τ が短いほど音の減衰が速くなる。本物のピアノの音声を聞くと、時間がたつにつれ低音が相対的に少し強くなっていったので、試みに f_1, f_2, f_3, f_4 について τ をそれぞれ 400, 380, 350, 300〔msec〕とした。

$$V_e(n) = \exp(-\frac{n}{f_s\tau}) \quad (n \text{ はサンプル番号、} f_s \text{ はサンプリング周波数}) \qquad \cdots\cdots(4)$$

実験結果

事前実験

録音



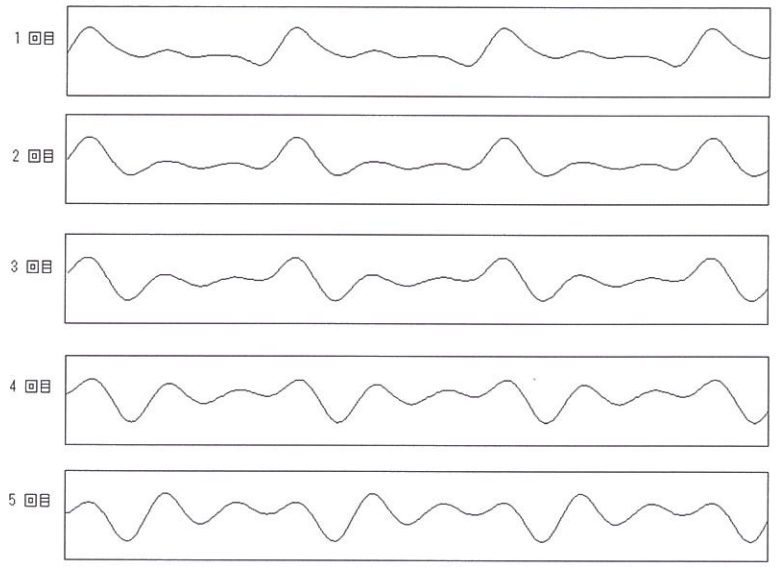
[図 8]　録音を繰り返した音の周波数成分

もとの音源は $a_1 : a_2 : a_3 : a_4 : a_5 = 10 : 6 : 8 : 2 : 2$ なる割合で sin 波を含む。

いま a_4 を一定として、スピーカーとマイクを通った回数による他の周波数帯の変化を、相対的に[図 8]で見る。

a_5 は急激に減衰し 3 回目でほぼ消滅し、a_1 も減衰している。a_2 はほぼ一定である。a_3 は増大してゆく。

a_3 の相対的増大は波形の凹凸を直接観察することでもわかる（[図 9]）。



[図 9]　録音を繰り返した音の波形

実験 1

実験 1-1

単調な繰り返しである S と S′ を足し合わせた W を、人間が S と S′ に分解することは、不可能であった。

実験 1-2

単なる音量変化としてしか知覚できなかった。また、フーリエ解析の精度を向上させるために [図 5]のような正弦波を「ハニング窓」として解析区間に掛けることがあるが、電子音では反対に精度が低下した。これにより音量変化が周波数成分に影響を与えることがわかった。

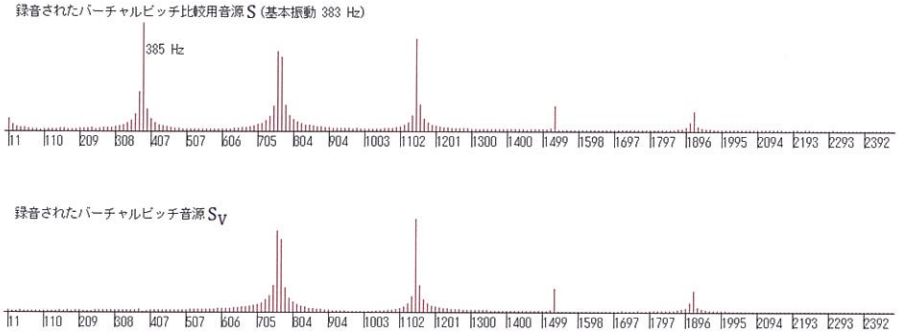
実験 1-3

全体の音量の変化が低減されたことで音色の時間変化を感じやすかったが、変化それ自体というよりむしろ、ある程度の楽器らしさを感じた。

実験 2

実験 2-1

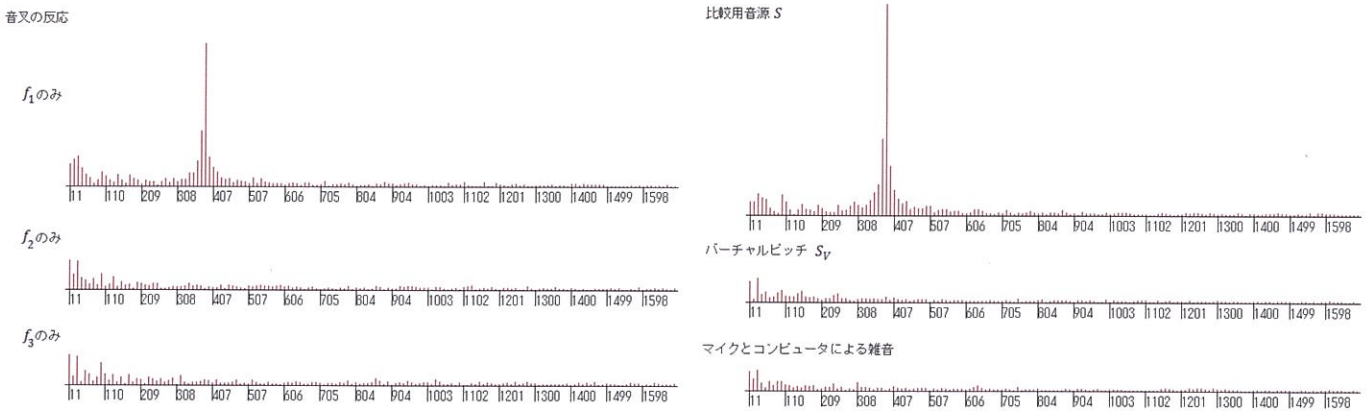
基本振動 f_1 を含む比較用の S と、f_1 を含まないバーチャルピッチ S_V をスピーカーから再生し、録音したものを周波数成分に分けると[図 10]のようになった。



[図 10]　録音された S と S_V のスペクトル

実験 2-2

音叉の共鳴をマイクにより録音した音声を 40 倍に拡大したのち周波数成分に分けると[図 11]のようになった。なお、十分に静かな部屋にマイクを置いて録音することで、マイクや録音用コンピュータ内部による雑音のスペクトルを得た（[図 11]右下）。

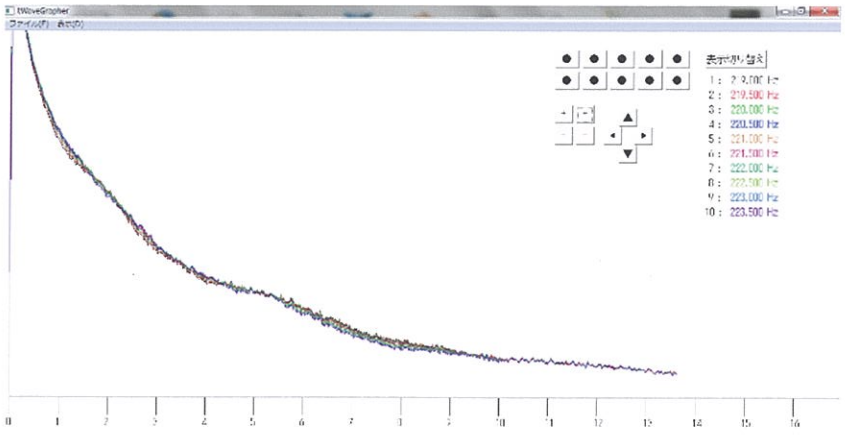


[図 11]　音叉に向け再生した音と音叉の応答（共鳴）の関係

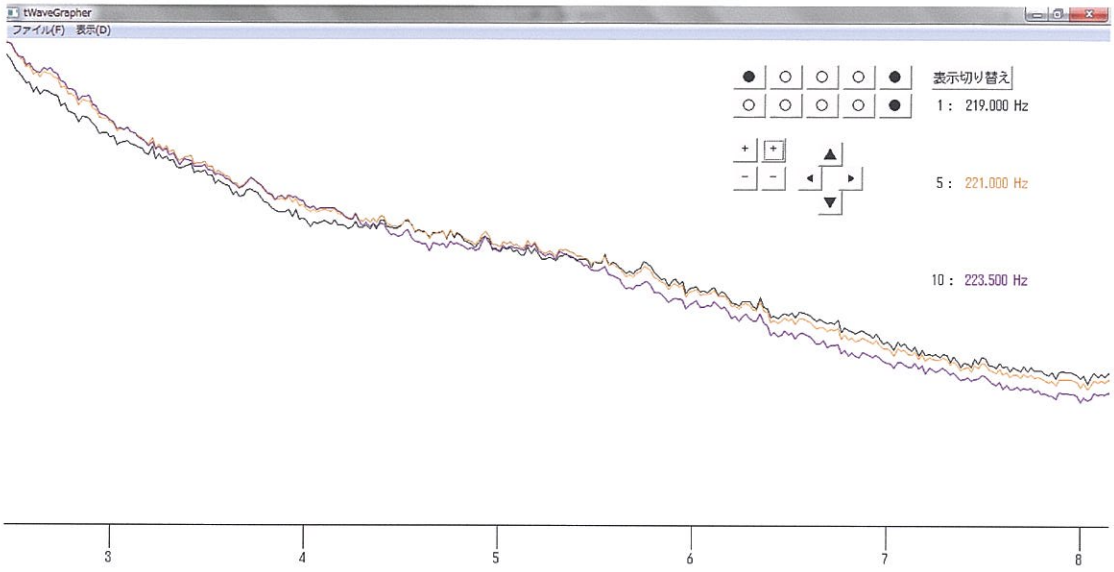
実験 3

実験 3−1

10 組のデータすべてをグラフにして重ねると[図 12]のようになった。また 1 番目の 219.000Hz, 5 番目の 221.000Hz, 10 番目の 223.500Hz だけを表示して一部を拡大すると、[図 13]のようになった。（[図 12]と[図 13]の横軸は時刻〔sec〕）



[図 12]　219.000 Hz ～ 223.500 Hz の振幅の時間変化が重ねられたグラフ



[図 13]　219.000 Hz（黒）, 221.000 Hz（茶）, 223.000 Hz（紫）の振幅の時間変化（一部拡大）

実験 3－2

ピアノの音にある程度近い音を合成することができた。

考察

実験 1

実験 1-1

この結果は考察のみによっても明らかである。

[図 14]を見るとわかるように、S と S′の各成分の振幅の値だけでなく各成分の sin 波の位相を考慮して足し合わせたものが W のスペクトルとなる。たとえば、f_3 成分のみ S と S′で逆位相（位相ずれ $\varphi = \pi$）としたため、W の f_3 成分の振幅は a_3 と a'_3 の差で表されているのがわかる。補足であるが、初期位相のずれ φ の同周波数の 2 つの sin 波を足すと同じ周波数の sin 波になり、ほかの周波数に影響しないことは(5)のように証明できる。

$$\begin{aligned} a_n \sin\omega t + a'_n \sin(\omega t + \varphi) &= (a_n + a'_n \cos\varphi)\sin\omega t + \sin\varphi \cos\omega t \\ &= \sqrt{(a_n + a'_n \cos\varphi)^2 + \sin^2\varphi}\,\sin(\omega t + \varphi') \end{aligned} \qquad \cdots\cdots(5)$$

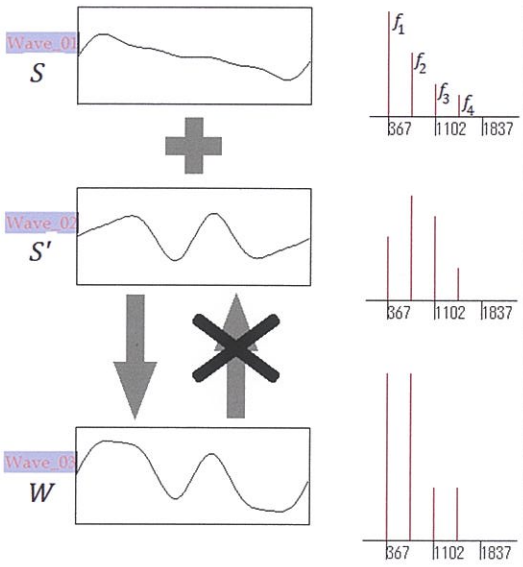
W のスペクトルを二つのスペクトルに分ける方法は、「S と S′」に限らず無限にあることは明白だ。このとき、W は不可分な一つの音声として知覚されると考えた。

【法則′】
ある音を複数のスペクトルに分ける方法が無限にあるとき、その音を人間はひとつの音として知覚する。

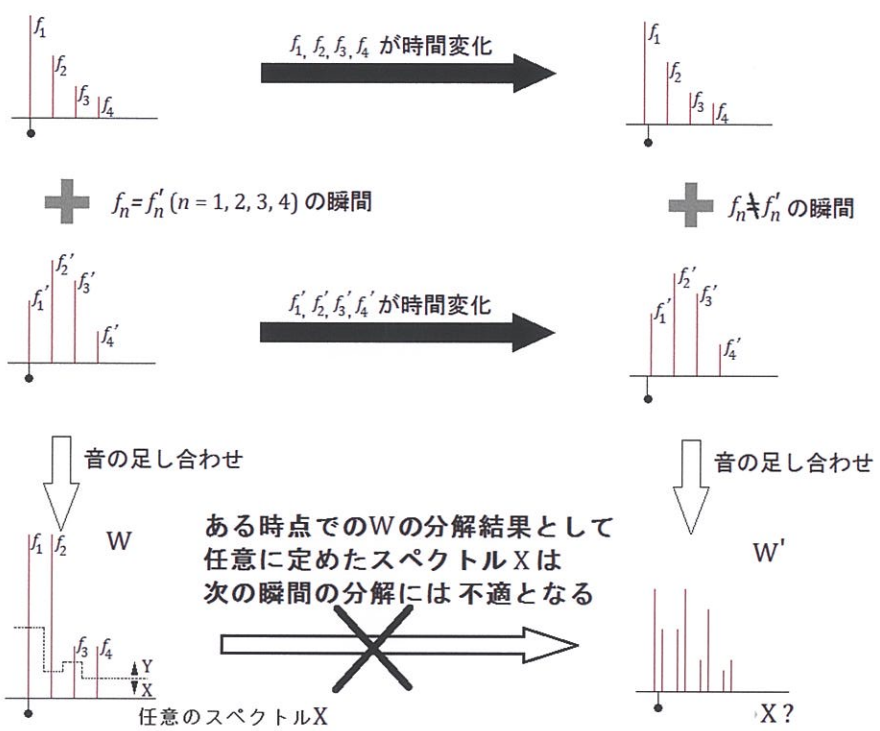
W を分解することができない理由を逆方向に考えることで、電子音でない二つの楽器が含まれる音声を二つのスペクトルに分解するときのことを考察した。【法則′】の条件を成立させているのは、S と S′のような電子音の特徴である「常に $f_n = f'_n$」なのではないだろうか。ここで、二つの音声のスペクトルの足し算の逆算をただ一通りに定めるには、上記【法則′】の条件部分を否定すればよい。すなわち、楽器音の特徴が「時刻によっては $f_n \neq f'_n$」であると仮定すると――つまり周波数が時間変化すると――、

ある時点における判断のみで定めたスペクトル（[図 15]左下のスペクトルを任意の二つに分けたうちの一つ）は次の瞬間（[図 15]右になったとき）には通用しなくなるだろう。このように、ある時点（[図 15]左）では無限に考えられる分解結果としての任意のスペクトルのうち、不適切なスペクトル X を選んでしまうと、X を抽出しようとする分解行為は時間軸上の連続性を失うのではないか。楽器の複合音を唯一の正しいスペクトルに分解することは、微細に時間変化する楽器の周波数群に対する分解行為が時間軸上で連続性を保つことと同値なのではないかと考えた。

補足であるが、楽器音の「時刻によっては$f_n \neq f'_n$」なる性質は以下によってもわかるだろう。オーケストラのように、楽器の数は多くなるほど全体の音の大きさは大きくなる。しかしそのすべての音が電子音のように変化のない音であったとしたらどのようなことが起こるか。いくつかの楽器との差異が考えられるが、その中の一つとして、音の時間軸上での位置、つまり位相によっては、音の数が増えても干渉によって**常に**一部が消しあい、全体の音は必ずしも大きくならないだろうということが挙げられる。楽器の数に音量が比例して聞こえるのは、波が時間変化をするからであろう。上記の、楽器音の特徴が「時刻によっては$f_n \neq f'_n$」である、との仮定はおそらく妥当であろう。



[図 14]　ただ一通りへの分解が不可能な場合



[図 15]　分解行為の時間的連続性について

上記の【法則'】に補足した以下を提案する。

【法則】
ある複合音を、**時間的な連続性を保った**複数のスペクトルに分ける方法が
- 無限にあるとき、ひとつの音として
- 唯一のとき、その音を楽器の複合音として

人間はその音を知覚する。

実験 1-2

S'の振幅が時間変化したからと言って、それが S と S'を区別できる糸口にはならない。W を S と S'へと分解することが不可能であって、W が単なる音量変化をしているようにしか知覚されないのは、分解の仕方が無限にあるためであると実験 1−1 と同様に考えることができる。

実験 1-3

W を S と S'へと分解できないのは言うまでもないが、全体の音量変化が低減したことにより、W は音色こそが時間変化する音声であると感じられるようになったと考えられる。W にはある程度の楽器らしさが感じられたから、音色の変化、つまり周波数成分の振幅の変化が楽器音を特徴づけるひとつの次元であると考えられる。

実験 2

実験 2-1

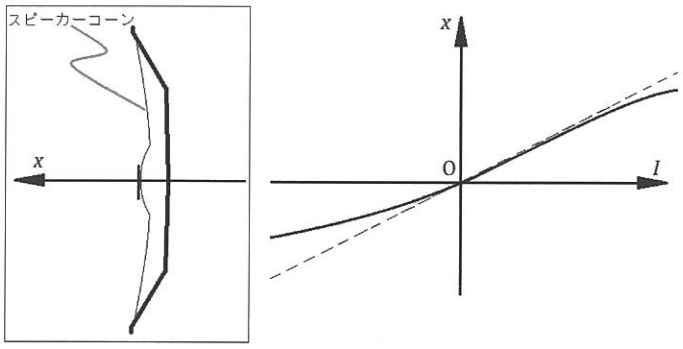
S_V を録音したものに明らかに f_1 は含まれていない。バーチャルピッチはスピーカーの性質によるものではないと断定する。

実験 2-2

高性能の録音機器を併用していないため、データを 40 倍に拡大したとき、主にコンピュータ内部に起因するノイズが特に 200 Hz 以下の低周波数帯に目立った。[図 11]においてノイズを解析したのはこのためである（ただし位相の考慮が必要であるから、他の解析結果からノイズの周波数成分を安易に減算することはさけ、併載するにとどめた）。

f_1, f_2, f_3 の sin 波を音叉に与えた時（[図 11]左の三つ）に、音叉の共鳴が観測されたのは f_1 のみであるということの確認ができた。f_1 を含む比較用音源 S は音叉を共鳴させたが、バーチャルピッチ S_V は共鳴させなかった。このことから、バーチャルピッチに聞こえる f_1 は聴覚に特有であって、f_1 を成分に持つ音声の物理的性質を持たないことが分かった。

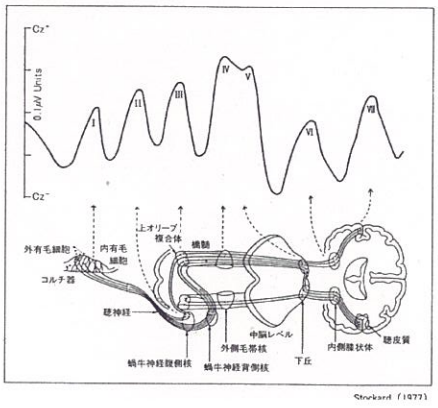
ではバーチャルピッチの発生要因は何であろうか。波形（[図 2]）を観察すると S と S_V の周期性は等しいから、sin 波に分解した各成分の強度だけではなくて、人間は空気振動の周期性によっても周波数をとらえているとすればバーチャルピッチが聞こえるはずである。空気振動を周期性自体で知覚することができる能力が人間に備わっていると考えられる。



[図 16]　流す電流とスピーカーコーンの変位（誇張されている）

一部の sin 波の位相のみを変えて合成した二つの音声を聞き比べると、成分の強度の比は同じであるにもかかわらず音色が微細に違った。このことからも聴覚のこの能力が裏付けられることを一旦考えた。しかしスピーカーコーンの変位 x が流した電流 I に厳密には比例せず非対称（[図 16]）である可能性があり、波形の違いが直接に成分の変化を引き起こすかもしれないため、これによる裏付けは避ける。同じ理由で、人間が成分の位相の違いにきわめて鈍感であることの真偽を一般的なスピーカーを使った実験で判断することは不可能であろう。

脳死判定にも用いられる聴性脳幹反応は、聴神経の応答速度を調べるものである。これによると、通常、蝸牛神経が音声を感覚するのは音の再生後 2 msec 程度だという。これは 441 Hz の音声では 0.882 周期分の長さにあたる。この速度で反応する能力をもつ聴覚関連器官は、空気振動の起伏（波形）を直接読み取るのに十分な能力を持つのではないかと推測する。



各ピークの起源	
I	蝸牛神経
II	橋延髄接合部
III	橋尾側(上オリーブ核)
IV	橋吻側
V	中脳(中脳下丘)
VI	内側膝状体
VII	聴皮質

[図 17]　聴性脳幹反応（岡山大学 http://www.okayama-u.ac.jp/user/hos/kensa/nou/ABR.htm より）

実験 3

実験 3-1

弦楽器の弦は、空気から抵抗を受けながらその振動を減衰させる。早く動くほど抵抗力が大きくなるという空気抵抗の性質と同様に考え、弦楽器の各成分の振幅は指数関数的に減衰すると予想していた。[図 12]から、おおまかな減衰は指数関数的であることが確かめられるが、それに緩く波打つ周期 3〜4sec の振動も加えられている。弦楽器を聞くとき、このような振動を聞くことはよくある。加えてさらに細かい振動が見られるが、解析に使用するデータ数によって DFT の結果が変わり、データ数が一波長分の整数倍（もっともそれを決定できるのは周波数が全く一定な場合に限るが）でない場合に結果の正確さが低下してしまうことに由来するものではないかと懸念した。しかし実験 3-2 で合成したピアノの音声を、解析に使用するデータ数を変えて結果を比較してみると、どの結果にも[図 12], [図 13]に見られる細かな振動は見られずグラフは滑らかな形状となった。このことから 10Hz 前後のこの振動は DFT の性質によるものではなく、楽器音に特有のものであると判断した。

もっとも重要なことだが、[図 13]の 3 本のグラフを追うと、時刻が 3 sec の周辺では 223.500 Hz, 221.000 Hz, 219.000 Hz の順に強いが、音量変化が起こることで 7 sec 以降では順が逆転していることがわかる。このことは、最も強い周波数が数 sec の周期で変化していることを表している。これも楽器音の特徴であると考えた。

結論

実験 1　複数の音声が重ねられた音声を、人間が元の組み合わせに分解することができるのは、各成分の周波数そのものが時間変化することによって決定されるただ一組の分解結果が導かれるときであると考えた。

実験 2　バーチャルピッチは人間に特有の現象で、その周波数の音声が持つ物理的性質を持たない。そして、人間は各周波数成分の強度だけでなく、空気振動の起伏とその周期性を読み取り、知覚に反映させていると考えた。

実験 3　ギターの音はその最も強い周波数の値を時間変化させ、音量は微細に振動しながら指数関数的に減衰する。a_n の比と音量変化というパラメータだけでも、ピアノを再現した音声を聞いてピアノとわかる程度の合成が可能だった。

マイクについて

考察を終えて、マイクの性質について疑問に思った。ダイナミックマイクは電磁誘導による起電力で動作するので、その点における振幅ではなく、その時間微分である空気の速度に（ほぼ）比例した信号が得られる（これが「速度型」と呼ばれる所以である）。つまり録音によってもとの信号を時間微分した信号が得られるはずであるから、[図 9]のように複数回マイクを通した波形を比べることに意味があるのか疑問に思った。一般的に「速度型」と言われていない、今回使用した ECM ではどうなるのか計算を試みた。

コンデンサーのダイアフラム（片方の極板のこと、音波により可動）の位置が時間 dt の間に変化して、もとは l だった極板間距離が dl だけ広がったとき、電気量が $Q_0 \to Q_1$ と変化したとする。

$$Q_0 = \varepsilon \frac{S}{l} V \quad \to \quad Q_1 = \varepsilon \frac{S}{l+dl} V$$

この間の電気量の変化量 dQ を、定かではないが $dl \ll l$ を仮定して表した（左）。これを時間微分して得られる電流（右）は

$$dQ = -\varepsilon VS\left(\frac{1}{l} - \frac{1}{l+dl}\right) \approx -\frac{\varepsilon VS}{l^2} dl \qquad \therefore I = \frac{dQ}{dt} = -\frac{\varepsilon VS}{l^2}\frac{dl}{dt} = -\frac{\varepsilon VS}{l^2} v$$

となり、I は変位 dl ではなく速度 v に比例するのではないか。

矩形波を録音すると、その微分波形であるインパルス信号にはならず、荒いがほぼ矩形波が得られた（[図 16]）。これは一見上式と矛盾するが、以下のように矩形波のフーリエ変換を考えることで解明されると考えた。

[図 18]　矩形波を録音したものの波形

$$(\text{矩形波}) = \sum_{n=1}^{\infty} \frac{1}{2n-1} \sin(2n-1)x$$

矩形波の微分は図形的に明らかなようにインパルス信号になり、そのままの右辺の微分では $\sin(2n-1)x$ の係数はすべて 1 である。しかし右辺において、高周波になるほどスピーカーとマイクを通すことで減衰するとすれば、微分した結果である録音信号がもとの矩形波のフーリエ変換に近くなることの辻褄が合う。以上により、一般的に言われていないようだが ECM を速度型マイクロフォンとみなしても差し支えないと考えた。

今後の展望

マイクと録音用コンピュータの総合的な性能の制約により、大きな音が出る楽器しか扱うことができなかった。しかし環境にはほかにも多種多様な音が溢れており、その一例として食器自動洗浄機が出す音は水の「じゃばじゃば」という音、水を循環させる装置の「ごおごお」という音、食器同士が当たる「カタカタ」という音などであると人間は分解できるが、そのような音声も機材（オーディオインターフェイス）があれば試みたい。

全体をとおして考察が難解であった。いまだに不十分な点があるかもしれない。また[図 15]に示したような思考に則り、楽器の複合音を分解する計算を実行する仕組みを構築したいが、かなり困難であると予想されるので、これは単なる希望にとどまっている。

今回の研究に限らず音声解析は、画像におけるパターン認識のように、たとえばこれから共生することになるだろうロボットたちが環境の音声の情報によって、われわれの生活環境において適切に行動するための技術として重要なものとなると考える。

参考資料

1. Kunio Kashino and Hidehiko Tanaka : A Sound Source Separation System with the Ability of Automatic Tone Modeling, 東京大学(1993)
2. 亀岡弘和 :「チュートリアル：非負値行列因子分解」, 東京大学大学院情報理工学系研究科 http://www.brl.ntt.co.jp/people/kameoka/publications/Kameoka2011MUS07.pdf
3. 岡山大学 :「聴性脳幹反応」http://www.okayama-u.ac.jp/user/hos/kensa/nou/ABR.htm

謝辞

実験で使用した音叉は、武蔵高校物理科研究室備品庫からお貸しいただいた。マイク（ECM）は秋葉原の千石電商様、アンプキットと 10W スピーカーは同秋月電商様から購入した。