バイオメディカル・ファジィ・システム学会誌 Vol.12, No.1 , pp 9−18 (2010)

[Original article]

(2010年4月23日 Accepted)

位相誤差を使用した音質評価指標

吉田 秀樹[1], 中野 正博[2], 行正 徹[3], 前田 康成[1], 横野 和也[1], 羽山 雄偉[1]

1）北見工業大学・情報システム工学科
2）産業医科大学・産業保健学部 3）産業医科大学・医学部・心理学

要約： 1オクターブに帯域制限した音響波形の極大値と極小値の情報があれば、元の波形を組み立てて情報を再現することができる。位相情報と振幅包絡はそれぞれ音源定位と音声情報の運び手の1つと関係付けられているので、計測した極値の位相と振幅方向の誤差のどちらが合成音の音質に重大な影響を及ぼすかについては関心が持たれるところである。そこで位相誤差と振幅誤差を独立に与えた合成音の主観評価をしたところ、2種類の合成音の波形と元の波形との二乗誤差は等しいにもかかわらず、位相誤差が有意に音質の劣化を招いていることを観察した。加えて極値を最小二乗推定することで位相誤差を4%未満に抑制すれば、合成音の音質が改善された。以上より、聴性認識は振幅より位相の検出に敏感にできており、位相誤差を算出すれば合成音の音質を見積もるための指標に利用できることが示唆された。

キーワード： 認識, 音響波形, 音声, 音, 合成

Cognitive Superiority of Phase Error to Amplitude Envelope in Sound

Hideki YOSHIDA[1], Masahiro NAKANO[2], Toru YUKIMASA[3],
Yasunari MAEDA[1], Kazuya YOKONO[1] and Yuui HAYAMA[1]

1) *Department of Computer Science, Kitami Institute of Technology*
2) *Dept. of Physics & Information Science, University of Occupational & Environmental Health*
3) *Department of Psychology, University of Occupational and Environmental Health*

Abstract: *It has been reported that the maximum and minimum data of the narrow-band (one octave) acoustic waveforms played a pivotal role in the sound reproduction by using the sinusoidal interpolation between two successive extrema. Although two indices of phase and quantization errors of the extrema have closely related to the perception of sound localization, and the form of amplitude envelope, respectively, which has regarded as a principal carrier of speech waveforms, the interaction between the two has remained unclear as to which is a critical factor in the auditory cognitive system. We produced two types of synthesized sounds, manipulating either time or amplitude of the extrema under condition that two sum squared errors between an unmodified sound and the two syntheses were mutually close. Scheffe's pair test has reported that the quality of synthesized 3-s speech and music with phase error was significantly lower than that of the corresponding sounds with quantization error, suggesting the cognitive superiority of phase error to noise in amplitude envelope. Besides, synthesized sounds with estimated extrema by using a least-square fit maintained almost the same quality of unprocessed sounds, when phase error was within the margin of four to nine percent, supporting the notion of an index, sensitive to the sound quality.*

Keywords: *cognition, acoustic waveform, speech, sound, synthesis*

Hideki YOSHIDA
Kouen-cho 168, Kitami, Hokkaidou, 090-8507, Japan
Phone: +81-157-24-9327, Fax: +81-157-24-9344, E-mail: hy@cs.kitami-it.ac.jp

1. はじめに

音によって運ばれる情報は神経の中を流れる電気信号に変換されて脳へと伝達される。音声も音楽も物音も全て聴神経（片耳当たり約３万本）の中を流れるスパイク波の時系列として表されており[1]、複雑な音響波形から情報を抽出し符号化する上では、進化の中で獲得された最適のアルゴリズム（算法）になっていると考えられる。音を電気信号に変える役割を担う蝸牛（直径約 9 mm）の内部は基底膜により仕切られており、基底膜上には振動センサーである有毛細胞（約１万５千個）が並んでいる。音が到来すると基底膜が振動し高い音は蝸牛の入り口近くで減衰するのに対して、低い音ほど強度を保ったまま基底膜上をより遠くまで（図５参照、約3 ms かけて約35 mm まで）進行する[2]。基底膜は機械的な周波数選別機として働いており、基底膜上の特定の場所に在る有毛細胞は特定の周波数に応答して電荷を誘発する。内有毛細胞の場合では基底膜が鼓室階方向（図５下向）に振れた時に最も高い電圧を生じる[3]。これは内有毛細胞が音響波形の山（あるいは谷）のどちらか一方の時刻と振幅を計測していることに他ならない。我々は蝸牛での振動−電気変換メカニズムに範を得て、音響波形の極大値と極小値を記録することで、逆に元の複合音の持つ複雑な高さ大きさ音色の合成（復元）ができることを報告した[4]。しかしながら、上述した内有毛細胞の検波特性[3]に従えば、極大値あるいは極小値のどちらか片方にしか対応していない情報を、脳内でどの様に補っているのかについては、知る限り何らかの考察や検討が加えられていない。そこで本研究では、従来から音色の特徴量の１つとされてきたスペクトル（周波数構造）ではなく、極値を音響波形の骨組みに据えて考察を行う。

音声の電算機による認識技術は、300-3,400 Hz 帯域（固定電話の通話帯域）に複数現れる声道での共振スペクトルを手掛かりにめざましい発展を遂げてきたのであるが[5]、一方で、音声信号は驚くほど狭い周波数帯域（1,500±150 Hz）に情報が集中していることも報告されている[6]。狭帯域のフィルタリングでは近接した周波数成分間で干渉が起き、振幅の強弱を交互に繰り返す所謂うなり様の時間波形が観察されるので、極値の取得が容易となる。同波形は基底膜の特定の場所で観られる振動に相当する。白色雑音のスペクトル包絡は平坦であるが、白色雑音を狭帯域の音声波の振幅包絡を使って変調（AM）してやると、雑音が音声信号として知覚できる様になるので[7]、音響心理学の立場からはスペクトル包絡よりもむしろ、振幅包絡が音声信号の担い手として重要視される根拠になっている。しかしながら、音声波の変調方式が AM 波に近いものであれば、振幅の持つ情報は伝搬経路上で重畳する様々な環境雑音（生活雑音）によって損なわれやすいので、頑健な音声コミュニケーションを実現する上では不利な要因ともなり得る。もし音響波形の厳密な時間形状の記録に聴覚情報としての価値が乏しいのであれば、従来の PCM(Pulse Code Modulation)録音技術を高分解能化していく進歩のあり方はヒトの認識特性から遠ざかるばかりか、合成音の品質（音声であれば明瞭性）を見積もるには、ヒトが試聴してみる以外に決定的な術が無いことも暗示している。

以上を踏まえて本研究の目的は、極値の時間軸方向への変位と、振幅方向への変位のどちらが音質に重大な影響を及ぼすのか、主観評価を実施することにある。狭帯域波形であっても音声を認識する上で充分な手掛かりが残されているので[6]、極値の時間情報に対応した有毛細胞の発火タイミングと、振幅情報に対応した発火強度のどちらが正確さを要する情報となっているのか導ける可能性がある。得られた観察結果は、原音を構成する極値と、合成音を構成する極値を比較することで、音質の劣化度を数値計算的に見積るための指標を与えるかも知れない。加えて、近接する極値の相対関係を統計的に観察すれば、聴覚が実現している巧みな情報抽出メカニズムを理解する糸口になるかも知れない。

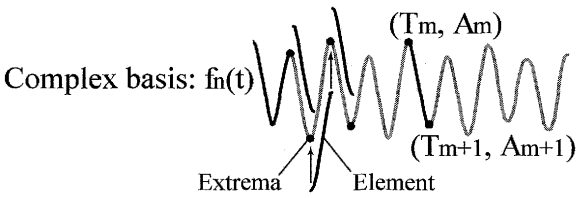
2. 音響波形の近似

帯域制限波は、離散フーリエ変換に従い、角速度を w_i として複素正弦波の部分和として記述できる。

$$fn(t)=\sum_{i=a}^{b}\exp(-jw_i t) \quad \text{where } j^2=-1 \qquad (1)$$

これは物理的には位相、振幅、周波数を違えた様々な正弦波の重ね合わせに他ならない。今、図１に示した様に、帯域制限波中の任意の極大値(T_m, A_m)から極小値(T_{m+1}, A_{m+1})まで、あるいは極小値から極大値までの一区間（$T_m < t < T_{m+1}$）を、1/2 周期長の正弦波を使用して近似する。提案する基底 $\phi_m(t)$は、図１上段に記した関数で定義でき、帯域制限波 $f_n(t)$は、重ね合わせと云うよりもむしろ、微小区間（エレメント）のつなぎ合わせで近似できる様になる。

$$\text{Basis: } \Phi_m(t) = \begin{cases} \frac{A_m - A_{m+1}}{2}\cos\frac{\pi(t-T_m)}{(T_{m+1}-T_m)} + \frac{A_m + A_{m+1}}{2} & \text{for } T_m < t < T_{m+1} \\ 0 \quad \text{otherwise} \end{cases} \quad \text{——(2)}$$



$$F(t) \fallingdotseq \sum_n f_n(t)$$

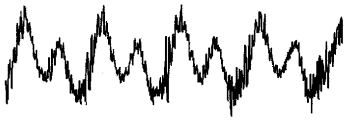


図１　基底 $\phi_m(t)$ と複合基底 $f_n(t)$

$$f_n(t) = \sum_m \Phi_m(t), \text{ for } m=1, 2, 3, \ldots M, \qquad (3)$$

合成した帯域制限波は基底をつなぎ合わせて造り出した複合基底であり、振幅と瞬間周波数を時々刻々と変化させる非常に複雑な構造をしている。そうして複合基底を重ね合わせることにより、観測された音響波形を近似することを提案する。

$$F(t) = \sum_n f_n(t) = \sum_n \sum_m \Phi_m(t), \text{ for } n=1, 2, 3, \ldots 10, \quad (4)$$

通過帯域幅を調整することで、エレメント（隣り合う極値間）は単調に増大あるいは減少する曲線になる。もし計測されたエレメントと正弦波との誤差を聴覚が聞き分けていなければ、以上の近似は音響心理の上では有効に働く（考察参照）[4, 8]。以下では実験的に通過帯域幅を１オクターブにとり、ヒトの可聴域を 10 チャンネルの複合基底で網羅させる。

Ch. 1, $f_1(t)$: 20-40 Hz　(5)
Ch. 2, $f_2(t)$: 40-80 Hz
Ch. 3, $f_3(t)$: 80-160 Hz
Ch. 4, $f_4(t)$: 160-320 Hz
Ch. 5, $f_5(t)$: 320-640 Hz
Ch. 6, $f_6(t)$: 640-1,280 Hz
Ch. 7, $f_7(t)$: 1,280-2,560 Hz
Ch. 8, $f_8(t)$: 2,560-5,120 Hz
Ch. 9, $f_9(t)$: 5,120-10,240 Hz
Ch. 10, $f_{10}(t)$: 10,240-20,480 Hz

チャンネルによっては単一の複合基底であっても、試聴して充分な意味をもたらすことが知られている[6]。複合基底を重ね合わせていくと、再生音の厚み、迫力、臨場感が増していく。音声を形造る主なスペクトル成分は、声帯での振動数(< 400 Hz)および声道での共振成分（< 4,000 Hz）であるので、本研究ではチャンネル３から８までの６個の複合基底を使用して音声を合成した。

３．　極値操作（実験１）

図２a に極値を変更することで波形を改変する様子を描写している。元の波形を破線で示し、変更後の極値は黒丸で示す。極値を振幅方向（左図）、あるいは時間方向（右図）に独立して変更し、これに伴い修正された波形を実線で示す。元の極値の振幅を A、山から山（あるいは谷から谷）までを一周期 D と考えて、極値の振幅方向への変量を a、時間方向への変量を d とおく。特定の極値にだけ大きな重みのついた修正が施されない様に、修正比 d/D と a/A を固定しておいて、正方向へ修正するか、負方向へ修正するかを乱数で決定する。ここで比較的振幅の小さな波形はエネルギー（音量）が小さく、聴覚に与える影響も少ないと予想される。各チャンネルでの帯域制限波の平均振幅を Aav とおき、Aav の 10 分の１未満の振幅であればノイズレベルとみなし、簡単のために極値の変更の対象から外した。時間方向と振幅方向への平均修正比 P と Q を統計的な指標に利用した。P と Q は極値数を N とおくと、以下の様に表される。

$$\text{位相誤差 : } P[\%] = \frac{100}{N}\Sigma\frac{d}{D} \qquad (6)$$

$$\text{振幅誤差 : } Q[\%] = \frac{100}{N}\Sigma\frac{a}{A} \qquad (7)$$

元の波形 **A** と、誤差を含んだ波形 **B** の物理的な差異は、二乗誤差を使用して定量化できる。下記の式は両波形 **A** と **B** の類似度を表している。

$$\text{合致率 : } MR[\%] = 100\left(1 - \frac{\Sigma(\mathbf{A}-\mathbf{B})^2}{\Sigma\mathbf{A}^2}\right) \qquad (8)$$

刺激には音声と音楽を使用し、CD-DA(Compact-Disc Digital Audio)規格に準拠したモノラル 16 ビット、44.1 kHz（非圧縮ディジタル録音方式の１つである.wav フォーマット）で録音したデータに、FIR(Finite Impulse Response)フィルターを適用して帯域制限波を作成した。タップ数はチャンネル２(80-160 Hz)が 2,207、チャンネル３(160-320 Hz)が 1,103、チャンネル４以降は 553

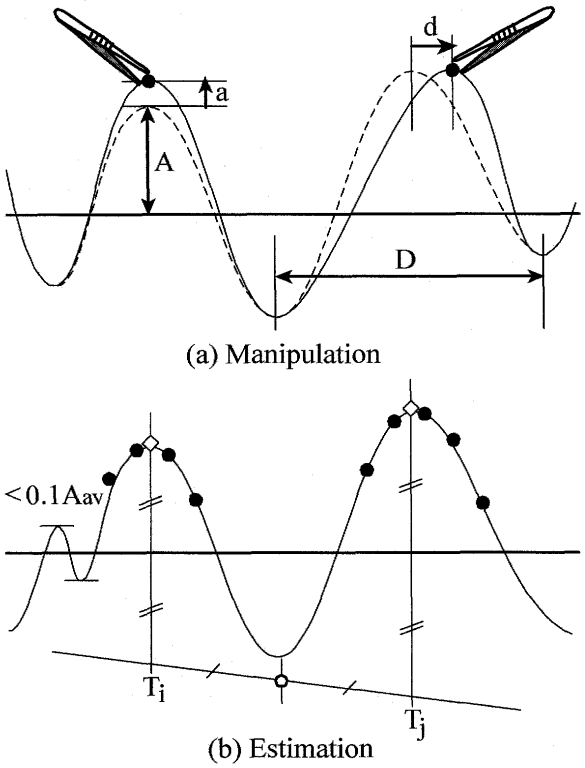


図２　極値の操作と推定　(a)黒丸は移動後の極値、実線は極値の移動に伴って修正された波形、破線は元の波形　(b) 黒丸はサンプリングされたデータ　菱形は推定された極値、白丸は推定された極値と時間軸対象にある２点 T_i と T_j の平均値

である。音声刺激としては、男声 1.5 秒と女声 1.5 秒から成る計３秒間の音声と４種類の合成音声を使用した。合成音声には、位相誤差を 4%与えた音声（以下 P4%と呼ぶ）と 7%与えた音声（P7%）が含まれる。原音と合成音声 P4%の間に生じる合致率 MRp が、同じく原音と振幅誤差を与えた合成音声の間に生じる合致率 MRq となるべく等しくなる様にして合成音声を作成し Q4%で表す。Q4%は安易に振幅誤差を 4%与えたと云う意味ではないので注意を要する。具体的には修正比 a/A を小刻みに変更しながら｜MRp−MRq｜＜１％となるまで反復計算を実施した。合成音声 P7%からも同様の手順で、適量の振幅誤差を与えた合成音声 Q7%を作成した。音声とは別に、クラッシック音楽を３秒間だけ切り出して、位相誤差を 7%および 10%与えた合成音 P7%と P10%、および同様にして応分の振幅誤差を与えた合成音 Q7%と Q10%を作成した。尚、音声刺激の周波数帯域を 80-5,120 Hz、チャンネル番号で云えばチャンネル３から８までとし、音楽刺激の帯域を 40-10,240 Hz、チャンネル２から９までとした。チャンネル１(20-40 Hz)の帯域の信号は通常観察されることが少なく、音源を特別に探して用意しなければならなくなるので本研究の観察対象からは外した。またチャンネル 10(10,240-20,480 Hz)の帯域に含まれる極値を計測するには、音楽の録音に良く使用されるサンプリング周波数 44.1 kHz では、後述の通り正確さに欠ける。広く普及した録音方式を活用する趣旨からチャンネル 10 も観察対象から外した。尚、当該チャンネル 10 は音声信号には影響が無く、音楽の再生時には臨場感が損なわれることがある。

４．　極値推定（実験２）

極値を計測するには、予めなるべく細かな時間幅（高いサンプリング周波数）で取得した時系列データから、時間軸に沿って高い値のデータと低い値のデータを交互に選び出してやれば良い。しかしながら、真の極値が存在する時刻とサンプリング周波数は無関係であるので、一般に真の極値は隣り合うデータの間に存在する可能性が高い。特に問題になるのが、採用したサンプリング周波数に比べて信号周波数が高い時であり、例えば本研究で採用した 44.1 kHz のサンプリング周波数では、チャンネル 10 帯域での極値の出現時刻の計測に大きな誤差が生じる。図２b は極値を推定する様子を描写している。黒丸はサンプリングされたデータの一部であり、これまではこの中から極大値を選び出していた。極大値を推定するには、頂上付近のデータを最小二乗法により放物線で近似した後に、放物線の頂点（菱形）を極大値として採用する[9]。極小値についても同様の手順が採用できるが、もっと簡易な方法は、接近した２個の極大値（T_i, A_i）と（T_j, A_j）について時間軸に対象な点（T_i, $-A_i$）と（T_j, $-A_j$）を求めておき白丸で示す平均値（$(T_i+T_j)/2$, $-(A_i+A_j)/2$）を採用する方法である。この様な推定値や平均値の使用が合成音の音質に影響を及ぼさないか、加えて事前に位相誤差を使って音質の劣化度を見積もれないかを調べた。刺激には実験１で使用した音声の原音 wav と合成音 P7%、および音楽の原音 wav と合成音 P10%を使用した。音声と音楽共に、放物線近似により極大値、極小値を推定して作成した合成音 X と、極大値のみ放物線から推定した後に極小値を平均値で代用した合成音 Y を作成した。

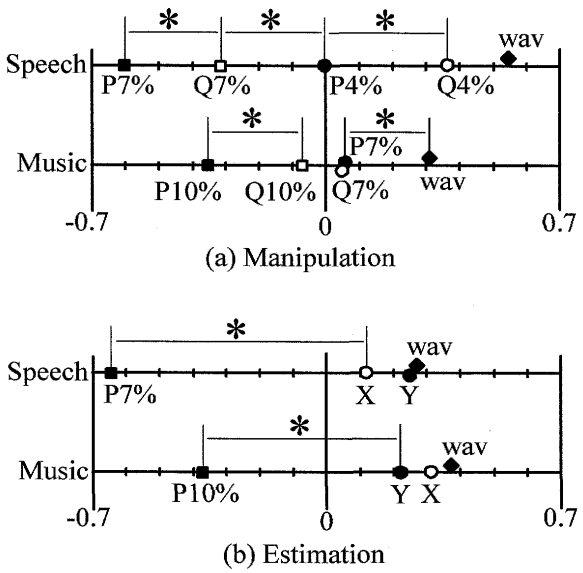


図３　主観評価値　(a)誤差を有した極値を使用した合成音　(b)推定した極値を使用した合成音　wav は原音、X は推定した極値を使用した合成音、Y は極大値を推定した後に極小値は平均値で代用した合成音、*p<.01

推定した極値と真の極値との間の誤差を計測した。予め厳密な極値を計測する必要があるので、サンプリング周波数１ MHz（CD-MA 規格外）で音声とクラシック音楽をそれぞれ１時間分のラジオ放送を録音した。FIR フィルターのタップ数はチャンネル２が 50,001、チャンネル３が 25,001、チャンネル４以降は 12,501 としてフィルタリングを実施して極値を計測した。同じ音声と音楽のデータを 44.1 kHz にダウンサンプリングしてフィルタリングの後、極値の推定を実施した。極小値は推定した極大値の平均値で代用した。

５．　主観評価

実験は 18 歳から 22 歳までの 50 人の被験者を対象にして、シェッフェの一対評価法を実施した。シェッフェの方法では、どちらがより大か判断させる一対評価法にカテゴリー判断を取り入れており、どちらがどれだけ好きかを評価点として回答させる。評価点は序数尺度であるが、評価点を間隔尺度に変換する手続きを省いて統計的検定が実施できる特徴がある。原音と合成音の組み合わせを継時的に２回呈示して聴き比べをさせて、被験者が判断した評価点を心理尺度上に表した。実験１の刺激音の組み合わせは、例えば刺激対 A と B の呈示順序を A から B とする場合と、逆順の B から A とする場合を数え上げることになる。加えて刺激の種類には音声と音楽を用いたので、組み合わせは併せて $2\times{}_5P_2=40$ 通りとなった。これをランダムな順序で被験者に提示して、音質の観点から４段階｛前者が鮮明、前者が僅かに鮮明、後者が僅かに鮮明、後者が鮮明｝で主観評価させた。一対評価法は刺激対の微妙な差異を判断する場合に適しているので、カテゴリーから中間値｛ほぼ同じ｝を外すことで、被験者が安易に中間値ばかりを連続して選択する可能性を排除した。同時にカテゴリー数を少なめの４段階としたことで、被験者には速やかな判断を促した。実験２の刺激音の組み合わせは正順と逆順および音声と音楽併せて $2\times{}_4P_2=24$ 通りであり、同様にして主観評価させた。刺激の音質は全て等価とした帰無仮説の下で ANOVA（分散分析法）を実施した。各刺激の評価点のクロス集計値は、心理尺度（数直線）上に換算して配置して刺激の順位を明らかにした。さらに任意の刺激間距離（各刺激間の関係性）Y は、スチューデント化された値 Ya を使用して有意水準 0.01 で下位検定することで明らかにした。

６．　結果

図３a に主観評価値を心理尺度（数直線）上に示す。音質は数値の大きな方が良質であり、菱形に wav と表した原音が最良と考えて良い。数値は零が中位で、負値をとると劣化（ざらつき）を感じる様になる。音楽の P7%、Q7%、Q10%を除き、評価値は心理尺度上に互いに充分な距離をおいて並ぶ様子が観察された。刺激要因に有意差が観られたので（音声 F(4, 744)=87.8、音楽 F(4, 744)=27.6）下位検定を実施したところ、音声では刺激 wav と Q4%組み合わせ以外の全ての組み合わせについて有意差（Ya=0.227 に対して Q4%－P4%間で Y=0.368、P4%－Q7%間で Y=0.310、Q7%－P7%間で Y=0.290）が確認された。音楽についても P7%、Q7%、Q10%間の組み合わせを除いて有意差（Ya=0.206 に対して wav－P7%間で Y=0.258、wav－Q7%間で Y=0.264、wav－Q10%間で Y=0.384、Q10%－P10%間で Y=0.280）が確認された。即ち、音声、音楽共に極値の誤差が増大すると音質が劣化しており、位相誤差（音声の P4% と P7%、音楽の P10%）の方が、応分の振幅誤差（音声の Q4% と Q7%、音楽の Q10%）よりも音質に大きな劣化が観察された。

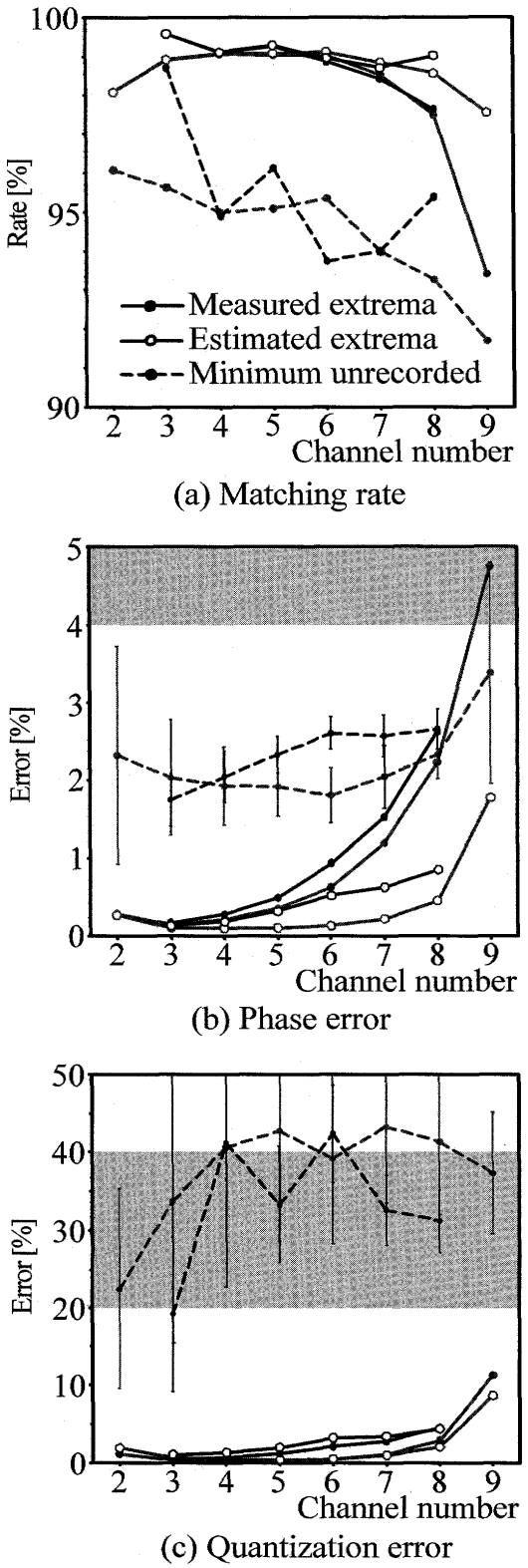


図4　誤差の周波数特性　(a)合致率　(b)位相誤差　(c)振幅誤差　黒色は音声（80-5,120 Hz）灰色は音楽（40-10,240 Hz）

図3b の心理尺度は、推定した極値を使用した合成音の音質を示している。音声、音楽共に、原音 wav と合成音 X と Y の主観評価値は接近した正値を取っているのに対して、位相誤差を含んだ極値から造り出した合成音 P7%と P10%は負値をとっている。刺激要因に有意差が観られたので（音声 $F(3, 397)=58.4$、音楽 $F(3, 397)=67.9$）下位検定を実施したところ、音声では wav、X、Y に対して P7%が有意に（$Ya=0.255$ に対して X－P7%間は $Y=0.760$）劣化していることが観察された。音楽でも wav、X、Y に対して P10%が有意に（$Ya=0.231$ に対して Y－P10%間は $Y=0.592$）劣化していることが観察された。音声、音楽共に wav、X、Y 間での有意差は観察されなかった。

7．　なぜ位相誤差に敏感であるのか？

先行研究では、位相誤差を 4%、9%および 14%の３種類、振幅誤差を 20%、40%および 60%の３種類与えた極値から刺激音を合成して主観評価を実施し、音質の劣化が位相誤差にして 4%から 9%の間、振幅誤差にして 20%、40%の間に始まることを報告した[4]。本研究では位相誤差による波形の変更の程度と振幅誤差による波形の変更の程度が、二乗誤差を介して等しくなる様にしたことで、最小数の刺激音の組み合わせで、位相誤差が振幅誤差より音質へ与える影響が大きいことを観察した。また音楽の P7% と Q7%でのみほぼ等しい主観評価値を示したことから、音声と音楽では認識特性に違いがあることも示唆された。これは脳が音声に対して高い分析性能を示した先行報告にも適っている[10]。本研究で定義した位相とは、隣り合う３個の極値を正弦波の一周期相当で結びこれを元の波形と見立てて、中央の極値の時間軸方向への進みあるいは遅れのことであるから注意を要す。任意のチャンネル間で波形全体を進めても遅らせても、音が滲んだり２つに分かれて聞こえたりする程度で、波形の微小区間を操作しない限り音質に重大な影響を与えることは難しい。この意味で、位相は同一音源から発せられた音波が両耳に到来する時間差（ITD: Interaural Time Difference）で定義される用法が一般的であり、音源定位の知覚と深く結びつけて議論されてきた。音源定位の中でも水平方向の推定の手掛かりは、約 1,500 Hz 以下の周波数成分で生じる ITD にあると報告されている[11]。純音では約 50 μs から ITD が知覚できるため、概算 $d/D=50\,\mu s \times 1{,}500$ Hz$=7.5\%$となり、これは音声刺激で明瞭にノイズ感を伴う位相誤差として観察した７%にほぼ合致している。すると狭帯域波形の極値を、

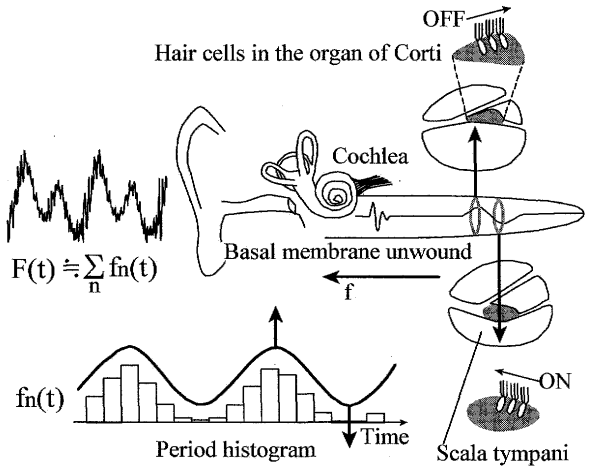


図５　内耳での振動-電気変換メカニズム

７%を超えてランダムに進めたり遅らせたりすれば、波形は現実には起こり得ない改変が施されたことになり、脳内での聴覚情報処理に混乱を招くことが予想される。他方、振幅誤差に関しては、音声波形を長さが50 ms以下（<20 Hz）の短い区間毎に分割して各区間について逆再生しても、音声全体としては明瞭度が維持される事が報告されている[12]。本実験の様に振幅をランダム改変する操作は、振幅包絡に高周波ノイズを重畳させるに留まり、20 Hz以下の振幅包絡が脳内で復元できる限りは、音声コミュニケーションに支障を来すことはないと云えよう。音声の空間伝送路で重畳する恐れの高い振幅誤差は、ヒトの聴性認識機構の中でも優先解決しなければならないノイズであり、直ちに脳内での情報処理に混乱を来す性格のものではないことが予想される。興味深い点は、性格の異なる２種類の誤差PとQのどちらが影響しても音質の変容振りは等しく、知覚されるのは耳障りなざらつき感であり、脳内で辛うじて認識可能な合成音の音質と云うのが、丁度チューニングの外れかかったラジオ音声の様な印象を得る。

従来の一定時間毎にサンプリングしたデータから真の極値の時刻と振幅を計測することは難しく、少なからず計測誤差を含めてしまう。実験１の結果を応用すれば、主観評価をする以前に、位相誤差を調べることで極値の推定が有効に働いているか見当をつけることができるかも知れない。図４a に、各チャンネルの帯域制限波と極値から合成した波形との合致率を、黒色で音声（チャンネル３から８まで）、灰色で音楽（チャンネル２から９まで）について重ねて示す。実線が計測した極値を使用した場合、実線に白丸が推定した極値を使用した場合、破線が極大値を推定した後で極小値を平均値で代用した場合である。図４b の位相誤差と図4c の振幅誤差でも記号の意味は同じとし、音声と音楽それぞれ１時間のデータを使用した。合致率でみれば、極小値に平均値を使用した場合に評価値がやや下がるものの、全例について 90%以上を示している。二乗誤差や相関係数を指標に使うと、例えば片方の波形の上下を反転するだけでも評価値を極端に下げることができるが、この操作により音質は変化しない。波形の寸分違わぬ正確さと音質とが必ずしも対応しないところに数値計算による音質評価の難しさがある。位相誤差（図４b）と振幅誤差（図 4c）はサンプリング周波数 1 MHz のデータから計測された極値を基準にして、サンプリング周波数44.1 kHz のデータから計測された極値、あるいは推定された極値について算出したものである。音声、音楽共に、上位のチャンネル（信号の周波数成分が高い）ほど位相誤差が大きいが、推定により位相誤差が改善（減少）していることがわかる。極小値を平均値で代用すると全チャンネルについて位相誤差を増大させるが、既に指摘した 4%の閾値未満に抑えられているので、音質への悪影響は問題にならないと予想される。これを振幅誤差でみれば、音声、音楽共に、極端に誤差は増大して、既に指摘した閾値20%から40%までに到達している。振幅誤差からだけでは音質へ与える影響の判断は難しいが、主観評価の結果(図３b)を良く反映しているのは位相誤差(図４b) の見積もりである。合致率より振幅誤差、振幅誤差より位相誤差が音質評価の上で本質的な指標となることが示唆される。

８．　なぜ極値から音は再生できるのか？

狭帯域波形の極値に着目する提案手法は、神経生理学の研究報告[13, 14]から支持される。図５に音波が聴神経の中を流れる電気信号に変換される様子を示す。入力音波は鼓膜と耳小骨の働きにより機械的な振動に変換されて、内耳にある蝸牛へと伝えられる。蝸牛内部の基底膜の振動には周波数特性が報告されており、特定の部位に配列した有毛細胞が特定の周波数の振動を検出する[15]。有毛細胞からは複数の不動毛が伸びており、有毛細胞で電荷を生み出すための引き金の役割をしている。音圧を受けて基底膜が鼓室階方向（図中下向）へ変位すると不動毛がたわみ有毛細胞を脱分極させ[3]、逆方向へ変位すると過分極させる。気圧の

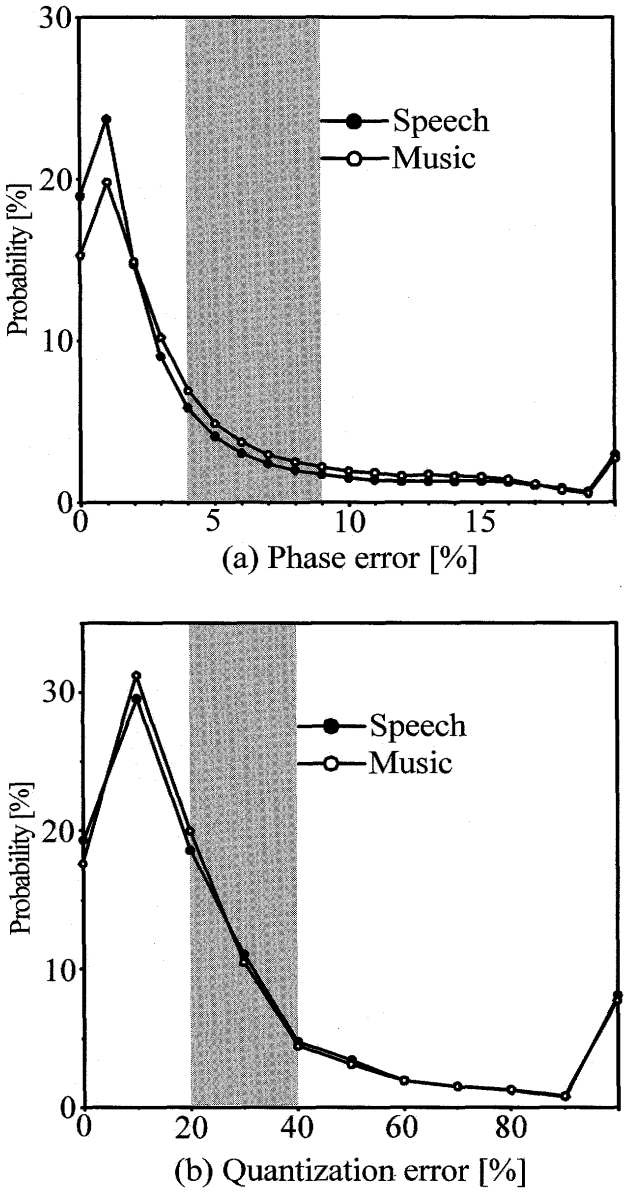


図６　極値の出現率　(a)位相誤差の分布　(b)振幅誤差の分布　いずれも計測した極小値の絶対値と隣接する極大値の平均値との誤差

疎と密であった音波が、基底膜の上下の動きで表され、図中のコルチ器(蝸牛管の断面図の中に着色した領域)の中に在る振動センサーで電荷に変換されたことになる。内有毛細胞の場合では過分極電位が小さいので、基底膜が片側に変位した時にだけ応答する。内有毛細胞につながるⅠ型聴神経では、入力波形の極大値あるいは極小値に同期してスパイク波が観察される。哺乳動物の聴神経での発火パターンの報告によれば、入力音波の周波数が４kHzまでに位相同期が観られる特徴がある[16, 17]。内有毛細胞より数の多い（不動毛の数も多い）外有毛細胞の場合では、脱分極による正電位は飽和しやすく、過分極によっても負電位が生成されるが、残念ながら外有毛細胞での誘発電位と、外有毛細胞につながるⅡ型聴神経での発火パターンとの関連については不明である[18]。提案した音声合成の手順と比べると、基底膜の振動が通過帯域を制限した波形に相当し、周期ヒストグラム中の発火数が極値の振幅、発火間隔が極大値から極大値（あるいは極小値から極小値）までの時間に相当する。脳神経での情報表現を説明する仮説は発火率コーディング[19]とテンポラル・スパイク・コーディング[20]の２種類に大別でき、いずれも神経発火が弱まっている期間には意味を与えていない。すると聴神経内部で表現された音波の特徴とは、狭帯域波形の極値情報のことであり、聴神経でのスパイク数が低下する信号休止期間は、謂わば聾区間になっている可能性が指摘される。従来から着目されてきたスペクトルにかかる情報処理は蝸牛から神経接続された蝸牛神経核[21]に始まり、スペクトル成分の抽出はさらに高次の情報処理に関わる下丘[22]から聴覚野で実現されていることが知られている。音響波形の再構成には、内有毛細胞の情報処理に習って極大値あるいは極小値のみの情報しか必要とされていないのか、あるいは外有毛細胞の誘発電位に習って極大値および極小値の双方の情報を必要としているのかであるが、これは次節での音響波形の特徴から考察をする。

９．　なぜ極小値は失われて良いのか？

極値が最小二乗法で推定できたのは、隣り合うデータに相関がみられるからである。音声波は自己回帰過程（Auto-Regression Process、全極モデル）[23]により記述でき、n個の計測されたデータからn+1番目のデータを推定する予測符号化技術として実用化されている。放物線を使った推定は予測器の設計を簡略化したものであるが、位相誤差の抑制には充分な効果がみられた[9]。同様にして極小値についても音響波形の構造上の規則性を導き出すことで、失われた極小値を尤もらしく復元するための説明を試みた。以下では精密な極値の計測を必要とするので、音声と音楽のそれぞれ１時間のデータに対してサンプリング周波数１MHzを適用した。図６ａに位相誤差の分布に対する、極小値の出現率を示す。位相誤差は、計測した極小値の出現時刻から、極小値を挟む極大値の中点までの時間を使用して算出した。音声（黒丸）、音楽（白丸）データ

共にグラフの形状は酷似しており、位相誤差が 4%、7%、9%までの極小値の割合は音声の場合でそれぞれ約 72%、82%、85%、音楽の場合でそれぞれ約 62%、77%、82%となりいずれも過半数を示した。極小値は位相誤差が大きくなるに従い出現率が減少するが、位相誤差が 20%を超える例外では出現率が増大に転じ、音声、音楽共に約 3%程度に達した。図６b に振幅誤差の分布に対する、極小値の出現率を示す。振幅誤差は、計測した極小値の振幅の絶対値と、極小値を挟む極大値の平均振幅の差をとり算出した。音声（黒丸）、音楽（白丸）データ共にグラフの形状はほぼ一致しており、振幅誤差が20%と40%までの極小値の割合は音声の場合でそれぞれ約 67%と 83%、音楽の場合でそれぞれ約 69%と 84%なり同じく過半数を示した。振幅誤差が 100%を超える例外では出現率が増大に転じ、音声、音楽共に約 8%程度に達した。出現率のピークが位相誤差で 1%、振幅誤差で 10%の所に観られたことから、波形の構造は過半数が対称的にできていることが示唆された。また音声と音楽には波形の極大値と極小値の関係に基づいた構造上の違いが観察されなかった。図４c で極小値を平均値で代用した時に振幅誤差が 20%から 40%程度に増大したのは、図６b の例外的に 100%を超える振幅誤差を有した極小値まで平均値で代用したからと考えられる。にもかかわらず主観評価では、原音 wav、推定値を使用した合成音 X、推定値と平均値を併用した合成音 Y との間に有意差は観察されていなかった。約８割の極小値については波形の構造から考えて平均値で代用することが認められるが、残り２割の例外については、蝸牛の外有毛細胞で極大値と極小値を計測した時の情報を補正信号として利用しているのかも知れない[3]。もし外有毛細胞に接続される II 型聴神経内の電気信号が、先行研究[18]にある通り、入力音波と無相関であるとすれば、２割の例外的な極小値は、脳内の情報復元機能[24]の働きで尤もらしく入力音波の情報が造り変えられて復元されている可能性も指摘できる。この場合、基底を半周期長の正弦波だけでなく、一周期長の正弦波と組み合わせて複合基底を組み立てるとした、修正音響構造モデルも視野に入れての追実験が望まれる。

１０. まとめ

音色自体に対応する物理量は存在せず、スペクトル、音の立ち上がり／立ち下がり／変化、タイミングにリズム、ノイズ等が音色の知覚に複雑に関わっていると解釈されてきた[25]。音色の構造は未解明であり、この意味でヒトの聴覚に頼らない音質評価は難しく、合成音の音質について見当がつけられるだけでも、ヒトの聴覚を最も簡素に模倣する糸口に成り得る。聴覚のメカニズムから考えてスペクトルの抽出はより高次の情報処理過程であるので、我々は振動−電気変換に当る蝸牛相当の音響構造モデルを提案した。当該モデルに基づいた合成音の音質は、狭帯域波形から許容誤差内で極値の時刻と振幅を計測することで維持される。波形の物理的な形状の合致度 MR よりも極値の振幅誤差 Q、振幅誤差よりも位相誤差 P が、聴覚情報処理により大きな影響を与えていることを主観評価により観察した。これは位相誤差を計測すれば合成音の音質を見積もるための一指標として利用できることを示唆している。許容される位相誤差は振幅誤差よりも小さく 4%から 9%程度であり[4]、最小二乗法により推定した極値の位相誤差が許容誤差内にあることを算出した。さらに推定した極値を使用した合成音が、見積もりの通り、原音に遜色の無い音質となっていることを主観評価により観察した。加えて、内有毛細胞の検波作用[3]で失われた極小値情報の復元に関しては、音響構造の微細な特徴に基づいて知る限り初めての考察も試みた。

参考文献

[1] Adrian, E. D. (1934): “The Basis of Sensation -The Action of the Sense Organs-”, London. Christophers.

[2] von Bekesy, G. (1960): “Experiments in Hearing” McGraw-Hill.

[3] Russel, I. J. and Sellick, P. M. (1978): “Intracellular Studies of Hair Cells in the Mammalian Cochlea”, Journal of Physiology, Vol.284, pp. 261-290.

[4] 吉田秀樹、角井健二、前田康成、藤原祥隆 (2008): “極値サンプリング技術と許容誤差- wavファイルからの情報抽出- ”, バイオメディカル・ファジイ・システム学会誌, Vol.10, No.2, pp. 123-131.

[5] Furui, S. (1986): “On the Role of Spectral Transition for Speech Perception”, J. Acoust. Soc. Amer., Vol.80, No.4, pp. 1016-1025.

[6] Warren, R. M., Riener, K. R., Bashford, J. A. Jr. and Brubaker, B. S. (1995): “Spectral Redundancy: Intelligibility of Sentences Heard through Narrow Spectral Slits”, Perception & Psychophysics Vol.57, pp. 175-182.

[7] Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. and

Ekelid, M. (1995): "Speech Recognition with Primarily Temporal Cues", Science, Vol.270, pp. 303-304.

[8] Yoshida, H., Kakui, K., Maeda, Y. and Fujiwara, Y. "Evaluation of the Synthesized Speech by Using Mismatch Negativity", International Journal of Biomedical Soft Computing and Human Sciences, under revision.

[9] Yoshida, H., Kakui, K., Maeda, Y. and Fujiwara, Y. "Least-Squares Estimation of the Extrema in the Narrow-Band Music Data", International Journal of Biomedical Soft Computing and Human Sciences, under revision.

[10] 吉田秀樹、角井健二、前田康成、藤原祥隆 (2008): "音響構造モデルと再構成技術-ノイズ音の認識特性-", バイオメディカル・ファジイ・システム学会誌, Vol.10, No.2, pp. 133-141.

[11] Makous, J. C. and Middlebrooks, J. C. (1990): "Two-dimensional Sound Localization by Human Listeners", J. Acoust. Soc. Amer., Vol.87, pp. 2188-2200.

[12] Saberi, K. & Perrott, D. R. (1999): "Cognitive Restoration of Reversed Speech", Nature, Vol.398, p. 760.

[13] Dallos, P., Billone, M. C., Durrant, J. D., Wang, C.-Y., and Raynor, S. (1972): "Cochlear Inner and Outer Hair Cells: Functional Differences", Science, Vol.177, pp. 356-358.

[14] Hudspeth, A. J. and Corey, D. P. (1977): "Sensitivity, Polarity and Conductance Change in the Response of Vertebrate Hair Cells to Controlled Mechanical Stimuli", PNAS, Vol.74, pp. 2407-2411.

[15] Greenwood, D. D. (1990): "A Cochlear Frequency -Position Function for Several Species- 29 Years Later", J. Acoust. Soc. Amer., Vol.87, pp. 2529-2605.

[16] Palmer, A. R. and Russel, I. J. (1986): "Phase-Locking in the Cochlear Nerve of the Guinea-Pig and It's Relation to the Receptor Potential of Inner Hair-Cells", Hearing Research, Vol.24, pp. 1-15.

[17] Attneave, F. and Olson, R. K. (1971): "Pitch as a Medium: A New Approach to Psychophysical Scaling", American Journal of Psychology, Vol.84, pp. 147-166.

[18] Robertson, D. (1984): "Horseradish Peroxides Injection of Physiologically Characterized Afferent and Efferent Neurons in the Guinea Pig Spiral Ganglion", Hearing Research, Vol.15, pp. 113-121.

[19] Robinson, D. A. (1975): "Oculomotor Control Signals" In: G. Lennerstrand and P. Bach-y-Rita (Eds.), "Basic Mechanisms of Ocular Motility and Their Clinical Implications", Oxford: Pergamon Press, pp. 337-374.

[20] Abeles, M., Bergman, H., Margalit, E., and Vaadia, E. (1993): "Spatiotemporal Firing Patterns in the Frontal Cortex of Behaving Monkeys", J. Neurophysiology, Vol.70, No.4, pp. 1629-1638.

[21] Young, E. D., Shofner, W. P., White, J. A., Robert, J. M., Voigt, H. F. (1988): "Response Properties of Cochlear Nucleus Neurons in Relationship to Physiological Mechanisms", Edelman, G. et al. (Eds), "Auditory Function", John Wiley and Sons, New York, pp 277-312.

[22] "The Mammalian Auditory Pathway" (1992): Popper, A. N. and Fay, R. R. (Eds.) Neurophysiology, Springer-Verlag.

[23] Itakura, F. and Saito, S. (1968): "Analysis Synthesis Telephony Based on the Maximum Likelihood Method", Reports of the 6th Int. Cong. Acoust., C-5-5.

[24] Warren, R. M. (1970): "Perceptual Restoration of Missing Speech Sounds", Science, Vol.167, pp. 392-393.

[25] "Hearing" (1995): Moore, B. C. J. (Ed), Academic Press Inc.

吉田秀樹（よしだひでき）
現職　北見工業大学工学部助教授
1991年 九州大学工学部電子工学科卒
1996年 博士（工学）（九州大学）
学会活動　BMFSA正会員
研究分野　医用生体工学

中野正博（なかのまさひろ）
現職　産業医科大学・産業保健学部
准教授・BMFSA学会長
1971年 九州大学理学部物理学科卒
1979年 理学博士（九州大学）
2009年 医学博士（産業医科大学）
学会活動　日本物理学会他多数
研究分野　原子核理論物理学・
統計学・情報科学

行正 徹（ゆきまさとおる）
現職　産業医科大学・医学部・心理学
准教授・精神保健指定医
1982年 京都大学理学部物理学科卒
1994年 産業医科大学医学部卒
2007年 医学博士（産業医科大学）
学会活動　BMFSA正会員他多数
研究分野　精神医学・原子核理論
物理学