■2群（画像・音・言語）－7編（音声認識と合成）

1章 音声基礎・分析・聴覚

（執筆者：）

■概要■

【本章の構成】

■2群 － 7編 － 1章

1-2 調音形態と調音運動の計測

（執筆者：党 建武）［2009年5月 受領］

人間は古くから自分自身の発話過程に興味を持ち，調音形態と調音運動を観測してきた．Hiltonは1836年に初めて腫瘍の手術による顔面欠陥の患者の発話運動を直接観測した[1]．それ以後，調音運動について侵襲的な観測が多数あった．Roentgenは1895年にX線を発見して，非侵襲的な方式で体内を観測できる写真記録法を確立した．その後，X線撮影により喉頭の動きや母音の計測が行われた[2-4]．X線による被爆の問題を回避するため，磁気センサや，超音波及びMRIなどの多様な計測技術や装置が開発された．本章では，音声研究に大きな影響を与えた，現在よく使われている技術をいくつか紹介する．

1-2-1 X線マイクロビーム

X線撮影において，放射線被爆の起因はX線の一部が人体に残されて吸収されたことである．もしX線に十分な透過力があれば，残らずに人体を透過し被爆量をかなり低減することができる．この発想に基づいて東京大学医学部音声言語医学研究施設において第1世代のX線マイクロビーム装置が開発された[5]．第2世代の装置が米国ウィスコンシン大学で開発され，同大学で，1987年に最初の実験が行われた．

(1) X線マイクロビーム装置の原理[6]

X線マイクロビーム装置の構成は，X線発生装置，X線検出装置，計算機システムからなる．X線発生装置は，通常のものと異なり，拡散X線を細いビーム状に形成して出力するものである．電子ビームを磁気により偏向させることで金属ターゲット上のX線発生位置を変え，ピンホールから放射されるX線ビームの方向を変えながら，観測領域を走差する．X線透過画像を得るため，走査に使われた偏向信号と検出されたX線強度に基づいて計算機により画像化する．運動物体の記録には，観測領域にある金属球（ペレットと称する）をX線ビームにより追跡する方式を用いる．計測を高速化するため，初期状態としてX線ビームにより観測領域の全面走査を行い，ペレットの初期位置を記録し，計測時には指標の周囲区域のみの走査によってペレットを確認したのちに,被観測体の運動開始と同時にペレットの逐次追跡を行いながら，ペレットの位置情報を実時間で記録する．

(2) X線マイクロビーム装置を用いる発話運動の観測

X線マイクロビーム装置による発話運動観測では，ペレットを舌上面にはりつけ運動指標として用いる．一般的なペレットの装着では，ペレットは舌表面上に4個または5個と，上唇，下唇と下顎門歯に1個ずつを設置する．発話時頭位変化による摂動を補正するために，座標変換の参照点として，上顎門歯と鼻背にペレットを2個設置する．通常，上顎咬合平面を水平軸とし，上顎門歯先端を原点とする座標系を基準座標として用いる．発話運動の計測には，ペレット初期位置の確認，開始音の出力，ペレットの追跡が自動的に行われる．発話記録の終了後に，口蓋の正中矢状面をペレットによりトレースし，口蓋形状を記録し，発話運動のペレットデータに合わせて処理が施され，基準座標系へ変換される．

(3) X線マイクロビームのデータベース

Westbury を中心とした米国ウィスコンシン大学の研究グループは，X 線マイクロビーム装置を用い音声生成データベースを作成した [7]．このデータベースには女性 26 名と男性 22 名から計測された音声と調音データが含まれる．音声資料は無意味の文を合わせて 116 文からなっている．舌，口唇，下顎などの発話運動は 8 個のペレットを用いて観測された．ペレットのデータは 146 Hz で標本化され，音声信号は 21739 Hz で標本化された．また，吐師らは同じ装置を用いて，『X 線マイクロビーム日本語データベース』を作成した [8]．このデータベースには，日本人成人 19 名による日本語発話データが収録されている．

1-2-2 磁気センサシステム

磁気センサシステム（Electromagnetic articulography, EMA）は，磁場において発話器官に貼り付けられたセンサコイルに起電電流の変化により，発話器官の動きを計測する装置である．その発想は熊本大学の園田により 1974 年に提案された [9]．その後，多数の研究者により改善され [10, 11]，1988 年に Carstens Medizinelektronik 社が初代の商用磁気センサシステム（AG100）を製造し，現在 5 次元 EMA（AG500）を製造・販売している．

(1) 磁気センサシステムの原理

EMA の原理は，センサコイルが貼り付けられた観測対象を磁場に置き，観測対象の動きに伴ってセンサコイルに電流を起電し，その電流の変化をセンサの動きに変換することである．装置の実装には，2 次元 EMA の場合 3 種類の送信コイル，5 次元 EMA の場合 6 種類の送信コイルを頭部周辺に配置して異なる周波数の磁界を生成して，調音器官に装着されたセンサコイルに電流を起電させる．その電流を周波数ごとに分離してコイルの位置に変換する．この方式では，マーカであるセンサコイルの位置や傾きは数百 Hz のレートで実時間で計測される．

(2) 磁気センサシステムによる計測

センサコイルは，一般的に顎，上唇，下唇，または軟口蓋にそれぞれ 1 個ずつ，舌面上に 3～4 個装着する．発話時の頭部の動きを補正するため，参照センサコイルは上顎や鼻などに装着される．EMA の計測では，同時に離散点を 10 個程度記録することが可能である．EMA による計測の特徴は，連続音声における口の動きがダイナミックに観測できることと，観測データが直接定量分析に用いられることが挙げられる．EMA は，調音運動の動的特性の観測にとって唯一の専用装置として現在広く用いられている．しかし，センサの間隔により観測点の数が限られる，コイルとワイヤが調音の妨げになる，などの問題がある．また，2 次元 EMA の場合，被験者の頭部の固定を要する．

(3) 磁気センサシステムによる調音データベース

5 次元 EMA 装置は安定性に多少問題があるため，大規模な計測にまだ使われていない．現有の調音データベースはほとんど 2 次元 EMA によるものである．そのなか，NTT の調音データベースは，成人男性話者 3 名により日本語 360 文を朗読したときの調音運動と音声データからなり，標本化周波数は調音データ 250 Hz で，音声信号 16 kHz である．また，エジン

バラ大学により作成されたデータベース MOCHA - TOMIT には，話者男女各 1 名に対して英語 460 文を朗読するときの調音データ，音声及びラリンゴグラフ（Electroglottograph）信号を収録した. 標本周波数は調音データ 500 Hz で, 音声信号とリンゴグラフ信号 16 kHz である.

1-2-3 磁気共鳴映像システム

1980 年代初期に臨床用として登場した磁気共鳴画像法（MRI）が，人体の断面を画像化する新しい方法として，様々な分野で応用されてきた．音声生成の研究への応用はその一例である．この観測法では輪切りの方向を自由に選べるので，垂直方向の断面（正中矢状面）も観測できるという利点がある. 多くの施設では 1.5 T－3.0 T の超伝導電磁石を用いた MRI が使われている.

(1) 磁気共鳴映像システムの原理

MRI の原理は，超伝導電磁石を用いて発生した強力な静磁場により生体にある水素原子スピンを磁化させた状態で急激な磁界の変化を与え，それらの原子から発せられる電磁波を記録して画像化することである．つまり，水素原子核は MRI のように強い磁場の中に置くと一斉に縦方向に整列する．更に，周波数 10～60 MHz の電波を照射することにより，水素原子核は静磁場方向から傾いて横向きに倒れ，その照射をやめれば徐々に元の状態に戻る．定常状態に戻るまでの過程で，それぞれの組織によって戻る速さが異なる．MRI では，各組織における戻り方の違いをパルスシーケンスのパラメータを工夫することにより画像化する．MRI に使われる照射電波はラジオ波の範囲であるため，被爆の心配はない.

(2) 磁気共鳴映像システムを用いる実験

音声研究の実験では，被験者に仰向きの姿勢で発話させ，発話器官の形状を記録して可視化する．MRI は基本的に静止画像の記録法で，3 次元の画像を得ることができるという利点があるが，映画のような連続的撮像は困難である．動画の同期撮像法と 3 次元動画撮像法は提案されたが，同じ動作の繰り返しが必要であるため，動作を再現するときの変動による画質の劣化がある．音声研究にとって一つの欠点は，MRI 装置から相当に大きな騒音が周期的に出るため，音声の同時記録が難しい点である．また，歯のような水分の少ない組織が MRI で画像化できないため，他の手法により歯列を補填する必要がなる．MRI による計測データについて，公開されているデータベースは少ない．2007 年 ATR-Promotions BAIC 事業部は日本語 5 母音の MRI データを公開した.

■参考文献

1) Bjork, L., “Velopharyngeal function in connected speech.,” Studies using tomography and cineradiography synchronized with speech spectrography., Stockholm:Appelbergs Borktrykeri, 1961.
2) Moeller, J. and J.F. Fischer, “Observation on the action of the cricothyroideus and throarytenoideus internus.,” Ann. Otol., Rhinol. Laryngol., vol.13, p.42-46, 1904.
3) Russell, G. O., “The Vowel: Its Physiological Mechanism as Shown by X-ray.,” Ohio State Univ. Press,1928.
4) Chiba, T. and M. Kajiyama, “The vowel: its nature and structure.,” Phonetic Society of Japan, 1941.
5) Fujimura, O., S. Kiritani, and H. Ishida, “Computer Controlled Radiography for Observation of Movements of Articulatory and Other Human Organs.,” Comput. Biol. Med., vol.3, p.371-384, 1973.
6) 本多清志，“X 線マイクロビームによる調音運動研究の動向.,” 音声研究，vol.2, no.2, p.8-18, 1998.

7) Westbury, J., “X-RAY MICROBEAM SPEECH PRODUCTION DATABASE USER'S HANDBOOK.,” Waisman Center, University of Wisconsin: Madison, USA., p.1-100, 1994.
8) 吐師道子, “X 線マイクロビーム日本語データベース,” 音声研究, vol. 4, no.2, p.31-35, 2000.
9) Sonoda, Y., “Observation of tongue movement employing magnetometer sensor.,” IEEE. Trans. Magn., vol. MAG-10, p.954-957, 1974.
10) Perkell, J., et al., “Electromagnetic midsagittal articulometer (EMMA) systems for transducing speech articulatory movements.,” J. Acoust. Soc. Am, vol.92, p.3078-3096, 1992.
11) Schönle, P., et al., “Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract.,” Brain Lang., vol.31, p.26-35, 1987.

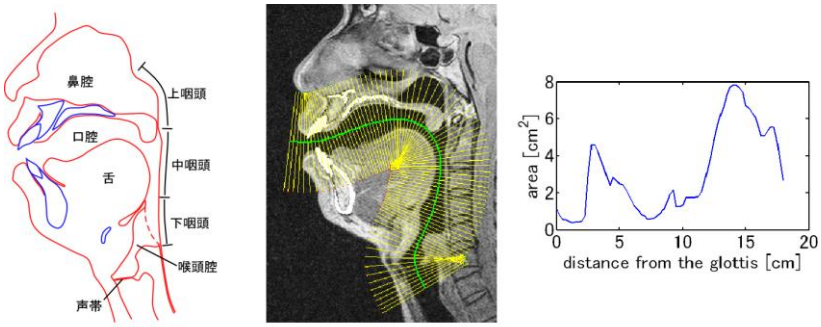
■2群 － 7編 － 1章

1-3 声道形状と音声スペクトル

（執筆者：竹本浩典）［2009年5月 受領］

声道とは声門上部にある腔で，喉頭腔，咽頭腔，口腔，鼻腔を指す（**図 1･1**(a)）．これらの腔は，舌などの運動器官によって形成されているため，その形状はダイナミックに変化する．声道の強い狭めによる乱流雑音や，声道の閉鎖と急激な開放による破裂音，そして声帯振動が，声道による音響的な修飾を受けることにより，種々の音韻が生成される．

母音では声帯振動が音源であるため，音声スペクトルは声道形状に固有の音響特性（伝達関数）によって特徴づけられる．また，鼻音化しない通常の母音では，軟口蓋が挙上して鼻腔は切り離されるため，声道には鼻腔は含まれない．しかし，鼻腔をのぞいても，声道形状は3次元的に極めて複雑であり，その形状から厳密な伝達関数を計算することは困難である．そこで，声道を断面積が変化する一本の管として近似し，平面波伝播を仮定できる低い周波数帯域で伝達関数を計算することが広く行われている（文献 1), 2), 3)）．なお，声門から口唇にいたる声道の長軸を声道中心線，これに対する断面積の変化を声道断面積関数と呼ぶ（図 1･1(b), (c)）．



(a) 声道　　(b) 声道中心線（緑）と断面積計測位置（黄）　　(c) 声道断面積関数

図 1･1　声道形状と声道断面積関数

剛壁と仮定した無損失の管の中では，空気の粘性を無視すると，以下の方程式が成り立つ．

$$\frac{\partial P(x,t)}{\partial x} = -\frac{\rho}{A(x)} \cdot \frac{\partial U(x,t)}{\partial t} \qquad (1 \cdot 1)$$

$$\frac{\partial U(x,t)}{\partial x} = -\frac{A(x)}{\rho c^2} \cdot \frac{\partial P(x,t)}{\partial t} \qquad (1 \cdot 2)$$

ここで，x は管に沿ってとった位置座標，t は時刻，$P(x, t)$は圧力，$U(x, t)$は体積速度，$A(x)$は位置 x における断面積，ρは空気の密度，c は音速である．これら二つの方程式は以下の一

つの方程式（Webster's horn equation）に結合される．

$$A(x)\frac{\partial}{\partial x}\left[\frac{1}{A(x)}\frac{\partial U(x,t)}{\partial x}\right]=\frac{1}{c^2}\cdot\frac{\partial^2 U(x,t)}{\partial t^2} \tag{1・3}$$

この式を適当な条件下で解くことも可能だが，ここでは，式(1･1)，(1･2)から等価回路モデルを導出する．(1･1), (1･2)で角周波数をω, 圧力と体積速度をそれぞれ $P(x, t)=P_0+p(x)e^{-j\omega t}$, $U(x, t)=U_0+u(x)e^{-j\omega t}$ とおくと，

$$\frac{dp(x)}{dx}=j\omega\frac{\rho}{A(x)}u(x) \tag{1・4}$$

$$\frac{du(x)}{dx}=j\omega\frac{A(x)}{\rho c^2}p(x) \tag{1・5}$$

となる．ここで，u を電流，p を電圧に対応させると伝送線路の方程式となり，その単位長さ当たりの直列インダクタンス $L(x)=\rho/A(x)$，並列キャパシタンス $C(x)=A(x)/\rho c^2$ となる．

図 1･2 は，例として声道を微小な長さ l を持つ n 個の円筒管の縦続接続で表現したときの伝送路（上）と，対応する四端子行列（下）である．ここで，L_1,…, L_n は各円筒管のインダクタンス，C_1, …, C_n は各円筒管のキャパシタンス，A_1, …, A_n は各円筒管の断面積，P_{in}，U_{in} は入力端の圧力と体積速度，P_{out}，U_{out} は出力端の圧力と体積速度，Z_r は放射インピーダンスである．

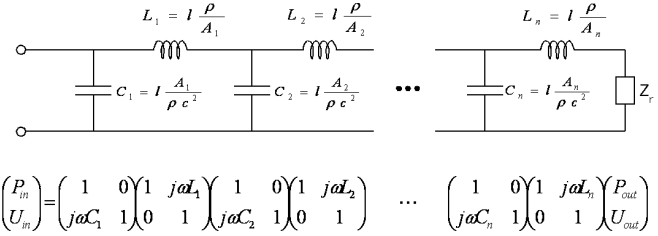


図 1・2　等価回路

行列の計算結果として四端子行列が以下のように表現されるとする．

$$\begin{pmatrix}P_{in}\\U_{in}\end{pmatrix}=\begin{pmatrix}A&B\\C&D\end{pmatrix}\begin{pmatrix}P_{out}\\U_{out}\end{pmatrix} \tag{1・6}$$

また，n 番目の円筒管の断面積を A_{out}，半径を r_{out}，波数を $k\,(=\omega/c)$とすると，放射インピーダンスは以下のように近似される [2)]．

$$Z_r=\frac{P_{out}}{U_{out}}\approx\frac{\rho c}{A_{out}}\cdot\left[\frac{(kr_{out})^2}{2}+j\frac{8kr_{out}}{3\pi}\right] \tag{1・7}$$

声門が完全に閉鎖していると仮定したとき，声道の体積速度に関する伝達関数は，

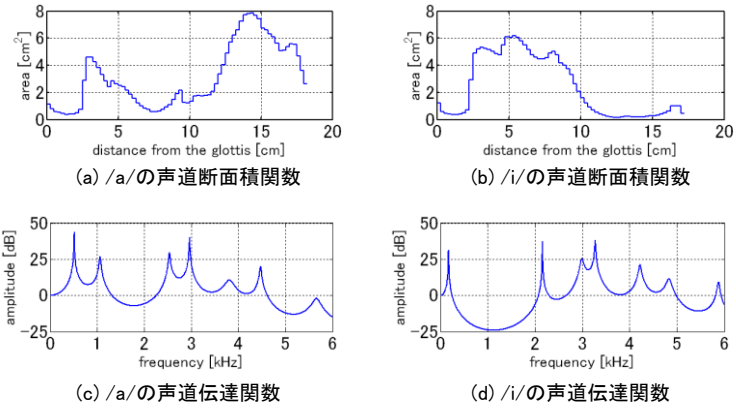
$$\frac{U_{out}}{U_{in}} = \frac{1}{CZ_r + D} \qquad (1・8)$$

である．

図 1･3(a), (b)は，それぞれ母音/a/と/i/の声道断面積関数である．断面積は声道中心線に沿って 0.25 cm ごとに計測したので，声道は長さ 0.25 cm の円筒管を縦続接続したものとして表現されている．母音/a/では声道の口腔部分が広く，咽頭腔部分が狭い．母音/i/ではこのパターンが逆になっている．

図 1･3(c), (d)は，それぞれ断面積関数から式(1･8)を使って計算した母音/a/と/i/の伝達関数である．なお，伝達関数のピークは音声スペクトルのホルマントに対応する．/a/の伝達関数では 1 番目のピークと 2 番目のピークが接近しているが，/i/では分離しているなど，それぞれの母音スペクトルの持つ特徴が現れている．

なお，伝達関数の計算において，声道壁の振動，熱交換，粘性抵抗による損失を考慮すると，1 番目のピーク周波数が上昇し，各ピークがなだらかになる（Q 値が下がる）ことが知られている．



(a) /a/の声道断面積関数　(b) /i/の声道断面積関数

(c) /a/の声道伝達関数　(d) /i/の声道伝達関数

図 1･3　声道断面積関数と伝達関数

■参考文献

1) J. L. Flanagan, “Speech analysis synthesis and perception,” Second, Expanded Edition, Springer, New York, 1972.
2) G. Fant, “Acoustic theory of speech production,” Mouton, The Hague, Paris, 1970.
3) B.H. Story, I.R. Titze, and E.A. Hoffman, “Vocal tract area functions from magnetic resonance imaging,” J. Acoust. Soc. Am., vol.100, no.1, pp.537-554, 1996.

■2群 － 7編 － 1章

1-4 聴覚経路

（執筆者：古川茂人）［2009年4月 受領］

1-4-1 末 梢

耳に到来した音は，外耳道を経て鼓膜を振動させ，三つの耳小骨の連鎖を通じてその振動が内耳へと伝えられる．内耳には，蝸牛と呼ばれる螺旋状の管（蝸牛）が存在する（**図 1･4** a,b）．蝸牛はその全長に渡り，ライスネル膜（音響的に透過）と基底膜によって三つの部分（前庭階，中央階，鼓室階）に仕切られ，内リンパ液（中央階）または外リンパ液（前庭階，鼓室階）で満たされている（図 1･4 c）．

耳小骨の振動によってリンパ液に変位が生じ，基底膜が振動する．基底膜は全長に渡って一様の性質を持っているわけではなく，正弦波入力に対する振幅包絡のピークは，高周波数では蝸牛基部近くに生じ，周波数が下がるに従って蝸牛先端部にシフトする．基底膜上の各点は，特定の周波数範囲に対してある程度選択的に振動する帯域通過フィルタ（聴覚フィルタ）として機能すると考えられ，基底膜は，中心周波数が規則的に変化する帯域フィルタ群とみなせる．

基底膜上には，その全長に沿って内有毛細胞と外有毛細胞が並ぶ（図 1･4 c,d）．細胞に付属する不動毛は，基底膜振動によって生ずる内リンパ液の流れに伴って変位する．不動毛がある方向に変位すると，その先端のイオンチャネルが開き，中央階の内リンパ液から K^+イオンが細胞内部に流入する．その結果，受容器電位が上昇（脱分極）する．受容器電位の変動は，入力波形の周期的変化に追従する交流成分と，それを積分したような直流成分が重畳されたかたちをとる．低周波刺激では交流成分が優勢であり，高周波数刺激では，直流成分が優勢となる．受容器電位の上昇と連動して，内有毛細胞とシナプス結合する聴神経の発火確率が上昇する．以上のメカニズムにより，単一聴神経の発火確率は，大雑把には，帯域フィルタ通過後に半波整流（イオンチャネルは，不動毛の一方向の変位に対してのみ開くため）した刺激波形を表現すると考えてよい．ただし，受容器電位の直流成分が優勢となる高周波刺激（> 1.5～4 kHz）に対しては，刺激の詳細な波形情報は失われ，時間包絡波形のみを反映する．

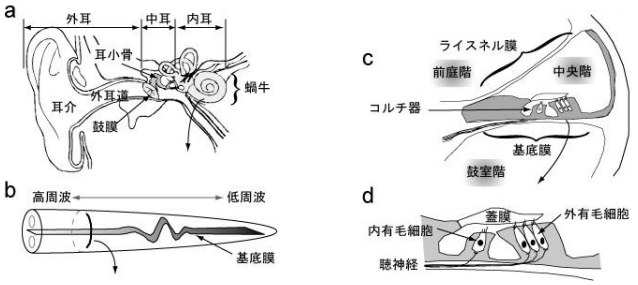


図 1･4　聴覚抹消系の概略．a: 外耳から内耳まで．b: 蝸牛の螺旋を引き伸ばした模式図．c: 蝸牛の断面の一部を拡大したもの．d： 基底膜上の有毛細胞部分を拡大したもの

外有毛細胞は，内有毛細胞と同様なメカニズムによって受容器電位が変化し，それに伴って細胞が伸縮する．その結果，基底膜の振動に影響を与える．つまり基底膜振動と外有毛細胞活動はフィードバック・ループを形成する．この機構は，微弱な信号に対する基底膜のゲインを上昇させるほか，基底膜での周波数選択性を先鋭化させる機能がある．感音性難聴者の多くに見られる聴力の劣化や周波数分解能の低下の背景には，外有毛細胞の機能損失があると考えられている．

1-4-2　中枢経路

求心性聴覚中枢経路の概略を**図 1･5** に示す．末梢から大脳皮質一次聴覚野（更に高次の領野でも）への求心性経路では，複数の神経核が介在する．そのうち主要な神経核において，各細胞は刺激の周波数にある程度選択的に反応し，その最適周波数に従って規則的に配置される傾向（周波数局在またはトノトピー構造）が見られる（トノトピー構造を持たない神経核もある）．中枢では，基底膜上の聴覚フィルタ群による周波数分析機構を基盤としながら，多段階かつ並列的な情報処理が脳幹で行われているといえる．

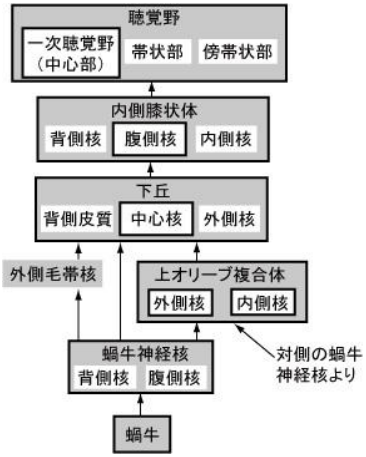


図 1・5　聴覚中枢系上向性経路の概略．片側半球の主要な神経核のみを示す．本文中で言及された部位は太線で囲ってある

主要な神経核について，その特徴をごく簡単に述べる．蝸牛神経核（cochlear nucleus, CN）は中枢の最初の中継核である．時間―周波数応答パターンによって CN 細胞は分類されているが，それらの機能的意味は十分に明らかになっていない．CN 細胞には，巨大な神経終末によって聴神経とシナプス結合しているものがある．これは，刺激音の時間波形を忠実に（あるいは強調して）伝達し，上位構造においてマイクロ秒レベルの両耳間時間差情報を処理するのに役立つ．上オリーブ複合体（superior olivary complex, SOC）のサブ構造である，上オリーブ内側核，外側核（medial/lateral superior olive, MSO/LSO）は，左右の CN から直接・間接

的な神経投射を受け，両耳に到達する刺激音の時間差及びレベル差（共に音源定位の手がかり）をそれぞれ処理すると考えられる．MSO 細胞の反応は，遅延線を介して両側から同時に到達する入力に対して強く応答するという神経回路モデル（Jeffress model）によって説明されることが多い．LSO 細胞は，一方の耳からは興奮性入力，他方から抑制性の入力を受けることで，両耳間レベル差に応じた反応を示す．下位の神経核からの求心性入力は，ほぼ必ず下丘中心核（central nucleus of inferior colliculus, ICc）を経由する．ここでは，最適周波数が等しい細胞がシート状に配列し，シートは最適周波数の順に分布している．シート上には様々な時間・周波数・レベル・両耳応答特性を持つ細胞が存在する．視床に位置する内側膝状体（medial geniculate body, MGB）は，求心性・遠心性接続によって聴覚野とフィードバック・ループを構成している．また，注意や覚醒状態などに関与する部位との密接な接続が知られている．一次聴覚野は，MGB 腹側核から求心性の投射を受け，トノトピー構造をもつ．その周辺の聴覚関連領域は，トノトピー構造の有無や，MGB などから投射様式の観点からいくつかの領野に分類されるが,動物種間でその定義は様々である.皮質領域間の水平接続や，求心性・遠心性の入出力接続が見られ，麻酔薬などによる細胞の活動特性への影響も顕著である．このため，高次の感覚・認知情報処理が聴覚野で行われていると考えられるが，具体的な機能は未知である．以上の求心性経路に加えて，皮質から末梢へと至る遠心性経路の存在や，聴覚系以外の脳内部位との関連も注目すべきであるが，ここでは割愛する．例としてあげる参考文献 1, 2)では，聴覚経路の概略に関する一歩進んだ解説が記載されている．

■参考文献

1) 平原達也，古川茂人，“聴覚の生理学,” 聴覚・触覚・前庭感覚，内川惠二（編），朝倉書店，pp.1-63, 2008.
2) J.F. Brugge, “An overview of central auditory processing,” The mammalian auditory pathway: Neurophysiology, eds A. N. Popper and R. R. Fay, New York: Springer-Verlag , pp.1-33, 1992.

■2群 － 7編 － 1章

1-5 聴覚脳活動

（執筆者：正木信夫）［2009年7月 受領］

音刺激により，蝸牛から大脳皮質の一次聴覚野までの聴覚伝導路や関連する部位で誘発される刺激関連電位（stimulus-related potential）を聴性誘発電位（AEP: auditory evoked potential）という [1]．また，一次聴覚野以後で行われる，認知処理機構を反映する電位は事象関連電位（ERP: event-related potential）と称される [2]．これらは脳波計測法（EEG: electroencephalography）により，その諸特性が明らかになってきた．近年は様々な脳活動計測技法が確立され，より高次の音情報処理に関連する脳活動部位の特定を目的とした研究が盛んである．本節ではAEP，ERP，及び人間を対象とする脳活動計測技術の概要を述べる．

1-5-1 聴性誘発電位（AEP）

AEPは，蝸電図（electrocochleogram）と頭頂部反応（vertex response）に分けられる．蝸電図は外耳道の鼓膜輪に置いた銀ボール電極または鼓膜から挿入される針電極と乳突部（耳介の後下方部）間の電位差で観測される．コルチ器や蝸牛神経が起源であり潜時は5 ms以下である．

頭頂部反応は頭頂部に置いた電極と耳たぶ，または乳突部の間の電位差で観測される．**図1･6**に代表的な頭頂部反応を示す [1, 3]．このうち聴性脳幹反応（ABR: auditory brainstem response）（図1･6中Ⅰ～Ⅶ）は蝸牛神経核から下丘に及ぶ広範囲の脳幹聴覚伝導路がその起源とみられている．続く聴性中間潜時反応（MLR: auditory middle latency response）（図1･6中 N_0, P_0, N_a, P_a, N_bなど）は聴皮質が起源と推測されるが，潜時の短い N_a-P_a-N_b は脳幹の関与も考えられている．また，頭頂部緩反応（SVR: slow vertex response）（図中 P_1,N_1,P_2,N_2）の起源は大脳の聴皮質とされる．この反応は覚醒時と睡眠時では異なる．図1･6は覚醒時の反応を示した．

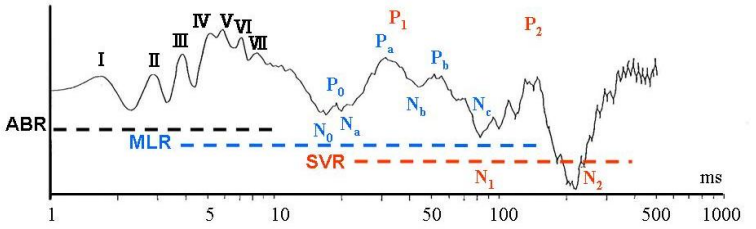


図1･6 主な聴性誘発電位の時間軸対数表示（文献1,3)よりカラー化して転載）

1-5-2 事象関連電位（ERP）

ERPの例をオドボール課題で観測される P_{300} で説明する．性質が異なる2種類の刺激（例えば周波数が異なるトーンバースト）をランダムに，ただしその一方の頻度を少なく提示する．低頻度刺激の数を数えるなどの作業を行わせ，低頻度刺激聴取時の脳波を加算平均する

と，高頻度刺激では出現しない陽性電位が潜時約 300 ms に見いだされる．これが ERP の一例 P_{300} である [2]．起源は大脳皮質連合野とされる [1]．

1-5-3　脳活動観測技術

現在，脳活動観測技術として確立している手法には，脳の神経活動に起因する電気生理的現象を直接的あるいは間接的にとらえる方法と，その神経活動に付随して生ずる血液動態や代謝の変化をとらえる方法とがある．以下その原理と特徴を概説する．

電気生理的現象に基づく脳活動観測技術

① 脳波計測法（EEG: electroencephalography）

耳たぶなどを基準として頭部に配置した電極との電位差を記録する [2]．電極配置は頭部に 19 個の電極を用いる 10/20 電極法が標準であるが，ヘッドキャップに数十の電極を搭載する機種もある．皿型の銀－塩化銀電極から導出された電気信号を増幅後，同期加算によりノイズを減弱させる．AEP・ERP の検討のほか，電源位置推定に基づく脳活動部位の検討も行われる．

② 脳磁計測法（MEG: magnetoencephalography）

脳の神経細胞内を流れる電流により発生する 100～1000 fT（フェムトテスラ=10^{-15} テスラ）の微小磁束密度を高感度磁気センサ，超伝導量子干渉計（SQUID: superconducting quantum interference device）で測定する [2]．最近では 200 チャネル以上のセンサを持つ装置が一般的である．EEG における AEP・ERP に対応する聴性誘発磁場（AEF: auditory evoked field）・事象関連磁場（ERF: event-related field）の検討のほか，磁場の等高線解析などに基づく電源（電流双極子（current dipole））の位置推定により脳活動部位の検討も行われる（**図 1･7**）．

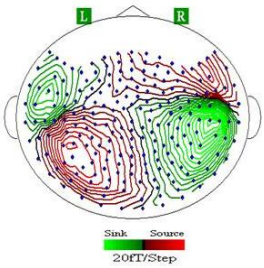


図 1･7　MEG による AEF 分析例（等高線表示）

血液動態あるいは代謝に基づく脳活動観測技術

① 機能的磁気共鳴画像法（fMRI: functional Magnetic Resonance Imaging）

神経細胞の電気的活動に伴う細胞周辺の毛細血管床の血流量変化を MRI の輝度変化としてとらえる [4]．提示刺激と同期して輝度が変化する部位を統計解析により mm の精度で推定する（**図 1･8**）．ただし刺激提示後，血液動態変化には数秒を要するため，EEG・MEG に比べて時間分解能は悪い．また装置自体が振動雑音を発するため，遮音性能のよいヘッドホン使用などの配慮が必要である．

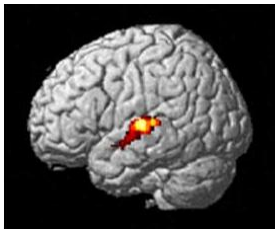


図 1・8　音刺激に対する fMRI の統計解析結果

② 陽電子断層画像法（PET: positron emission tomography）

脳に送り込まれる血液中に ^{15}O, ^{11}C, ^{18}F などの半減期の短いポジトロン放射核種で標識した標識薬剤を注射し，脳活動に伴う脳血流量，脳血液量，脳酸素摂取量，脳酸素消費量，脳ブドウ糖消費量等の分布を画像化する [2].

③ 近赤外線分光法（NIRS: near infrared spectroscopy）

複数波長の近赤外線光を光ファイバで頭皮上から頭蓋内に投入し，波長ごとの吸収率の違いから脳表近傍の血液中の酸化ヘモグロビン・還元ヘモグロビンの量を推定し画像化する [5]. 他の手法に比べ被験者の拘束性は少ないため，幼児の実験などにも用いられる.

■参考文献

1) 真鍋敏毅,“第 6 章　聴性誘発反応の概略,”“聴性脳幹反応［その基礎と臨床］”鈴木篤郎（監修），船坂宗太郎・大西信治郎（編集），メジカルビュー社，pp.82-87，1985.
2) 柴崎浩, 米倉義晴,“脳のイメージング－脳のはたらきはどこまで画像化できるか,”共立出版，1994.
3) 市川銀一郎, 河村正三 他,“聴性誘発反応の対数時間軸表示法,”Audiology Jpn, vol.26, pp.735-739, 1983.
4) S.A. Huettel, A.W.Song and G.McCarthy, “Functional Magnetic Resonance Imaging,” Sinauer Associates, 2003.
5) 星詳子，田村守，“近赤外線による脳代謝測定,”神経研究の進歩，vol.38, no.2, pp.301-308, 1994.

■2群 － 7編 － 1章

1-6 聴覚モデルの基礎と応用

（執筆者：鵜木祐史）［2009年5月 受領］

聴覚モデルとは，私たちの「耳」，すなわち聴覚系を機能的あるいは物理的に模擬したものであり，そのメカニズムを十分に説明できるものでなければならない．聴覚経路（1-4 節）で説明されたように，聴覚系といってもその定義の範囲は広く，研究分野によっても異なるが，一般には聴覚末梢系（蝸牛）までを指すものと考えられる．ここでは，聴覚末梢系を模擬したものを聴覚モデル，聴覚経路による分類をせずに知覚的側面（例えばピッチやラウドネス）から模擬したものを聴知覚モデルと呼び，切り分けることにする．本節では，前者の聴覚末梢系を模擬した聴覚モデルの基礎と応用を説明する．

聴覚末梢系における信号処理は，**図 1･9** に示すようなブロックダイアグラムで分類される．モデル化の対象は，外界から順番に，外耳（耳介・外耳道），中耳（ツチ・キヌタ・アブミ骨），内耳（蝸牛）及び末梢神経（1 次聴神経）となる．外耳は主に高域強調の機能を持ち，空間知覚で重要であるといわれている．中耳は，空気の圧力変化を効率良く蝸牛内の振動変化（リンパ液の圧力変動に伴う基底膜の振動）として伝搬させるための音響インピーダンス整合の機能を持つ．蝸牛（基底膜，内有毛・外有毛細胞）は音成分の周波数分解，いわゆるスペクトル分析機能を持ち，その分析は音の周波数や音圧レベルに依存し，非線形で能動的な振る舞いをする．末梢神経は，それらを聴神経インパルス列として符号化し，中枢から高次（求心性神経）へと伝える役目を担う．聴覚末梢系には，この一連の流れとは逆向きに，蝸牛の非線形性として知られる耳音響放射や，遠心性の情報による内耳・末梢神経への影響並びに耳の保護の役割を果たす耳小骨筋反射が生ずる．これらの現象を含めて模擬されたモデルがより精緻な聴覚モデルとなるが，本節では，図 1･9 の右向き（外界から求心性神経へ）の流れを模擬したフロントエンドの聴覚モデルを説明する．

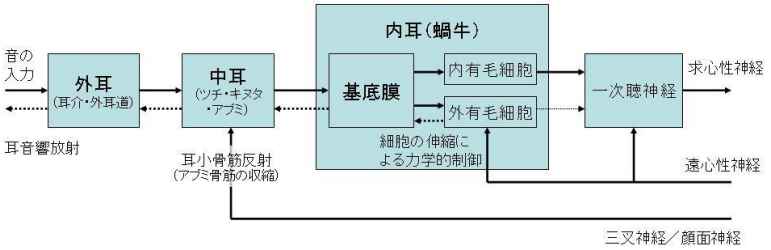


図 1・9　聴覚末梢系のブロックダイアグラム

1-6-1　聴覚フィルタ

聴覚末梢系で行われている音のスペクトル分析は，工学的にいえば音の入力に依存したフィルタ特性（フィルタの Q 値，帯域幅，対称・非対称な形状変化）を持つ帯域通過フィルタ群の処理と解釈できる．この帯域フィルタは聴覚フィルタと呼ばれ，これが周波数方向に連

続して並んだフィルタ群は，聴覚フィルタバンクと呼ばれる（後述するが，最初の定義は心理物理的意味づけからきている）[1]．一般に，聴覚末梢系の周波数選択性（複合音を正弦波に分解する能力）は聴覚フィルタバンクを利用して説明され，心理物理的アプローチからは心理物理的同調特性あるいは同時・非同時マスキング特性を調べることで得られる．生理学的アプローチからは，音信号に対する基底膜振動の特性（レーザドップラー法を利用して測定）や，その後段にある聴神経繊維の発火特性から直接得られる生理学的同調特性や等強度関数などを調べることで得られる．しかし，ヒトに対して，聴神経発火パターンといった聴神経線維の特性を生理学的手法で直接調べることができないため，代替案として，ネコやモルモットなどを対象とした動物実験から聴覚フィルタに関する生理学的知見を得ている．

ヒトの聴覚フィルタに関する旧来の推定法には，次に述べる二つの心理物理学的手法がある．一つは低レベル（例えば感覚レベルで 10 dB）の正弦波を信号音，もう一つの正弦波（あるいは狭帯域雑音）をマスカーとしてソーイングしてマスキング閾値を求めることで，心理物理的同調特性を調べる手法である．これは生理学的同調特性と高い相関を持つが，二つの正弦波の周波数の関係によってうなり現象が生ずるため，聴覚フィルタの先端部を正確に推定することができない．もう一つの手法は，マスキングのパワースペクトルモデルを利用する方法である．最初の有名な実験として，Fletcher による狭帯域雑音を利用した同時マスキング実験[1]がある．この実験をきっかけに，聴覚フィルタ（一種の帯域通過フィルタ）が仮定され，信号音をマスクする雑音成分が聴覚フィルタ内の周波数成分に限られると考えられるようになった．ここで得られたフィルタ帯域幅は臨界帯域（CB）と呼ばれ，更に CB を幅 1 として周波数軸を変形したものは，Bark 軸と呼ばれた[2]．しかし，この実験には離調聴取の問題（信号を検知するときにその周波数とフィルタの中心周波数を一致させて聴いているわけではない）[3]が残っている．これを避けるために，Patterson & Moore によってノッチ雑音マスキング法[4]が確立され，以後，聴覚フィルタ形状が非対称であることが明らかになった．

現在最も主流で有効な聴覚フィルタの推定法は，ノッチ雑音マスキング法[4]を利用してマスキング閾値を求め，マスキングのパワースペクトルモデルを仮定してフィルタ形状を推定することである（**注**：ELC 補正や中耳の伝達特性の補正をすることで外耳・中耳の特性を取り除いて推定している）[5,10]．代表的な方法として，PolyFit 法[5]（マスキング閾値データへの聴覚フィルタを表現するパラメトリック関数の最小自乗適合）がある．この実験結果に基づいて得られたフィルタの帯域幅は，等価矩形帯域幅（ERB）と呼ばれ，また ERB を幅 1 として周波数軸を変形したものは ERB-number と呼ばれる[5]（**注**：健聴者と難聴者のフィルタ帯域幅を区別するために健聴者の ERB を ERB_N，ERB-number を ERB_N-number と定義している[6]）．最近では，蝸牛の圧縮特性（増幅特性）や他の非線形作用（例えば二音抑圧）の心理物理的測定法（例えば，順向性マスキングといった非同時マスキング実験）とフィルタ形状の検討も行われ始めている[11,12]．

1-6-2　聴覚モデル

心理物理のアプローチから提案された聴覚フィルタ関数として，roex（rounded-exponential）フィルタ[3,4,5]，Gammatone フィルタ[7]，Gammachirp フィルタ[8]が知られている．roex フィルタは周波数領域で独立に定義されるため，フィルタ形状（特に頂上／裾野の形状）を非対称に表現できるが，インパルス応答を持たないことから，時間領域のフィルタバンク処理を実

現できない．Gammatone フィルタはインパルス応答関数として定義されるため，wavelet 変換や IIR フィルタとして時間領域のフィルタバンクを構成できるが，対称なフィルタ形状しか表現できない．また，この処理は線形で受動的である．Gammachirp フィルタは Gammatone フィルタにチャープを持たせることで，フィルタ形状の非対称性を表現することができ，また心理物理データだけでなく生理学的知見とも良い整合性を持つ．更に，これは蝙牛で観測される圧縮特性を模擬した圧縮型 Gammachirp フィルタ [9] に拡張されている．**図 1･10** は，ヒトの大規模な心理物理データに適合して得られた聴覚フィルタの特性 [10]を示す．ここでは，音圧レベル依存で，能動的に非線形な振る舞い（圧縮特性）を示す特性が得られている．最近では，瞬時的な圧縮特性や二音抑圧を定性的に説明可能な動的な圧縮型 Gammachirp フィルタバンク [11]も報告されている．心理物理から得られた聴覚フィルタのほとんどは，同時マスキング実験の結果に基づくものであり，順向性マスキング実験などに基づいた検討 [12] も出始めているが，動的な特性を直接組み込んだ聴覚フィルタバンクはまだ提案されていない．

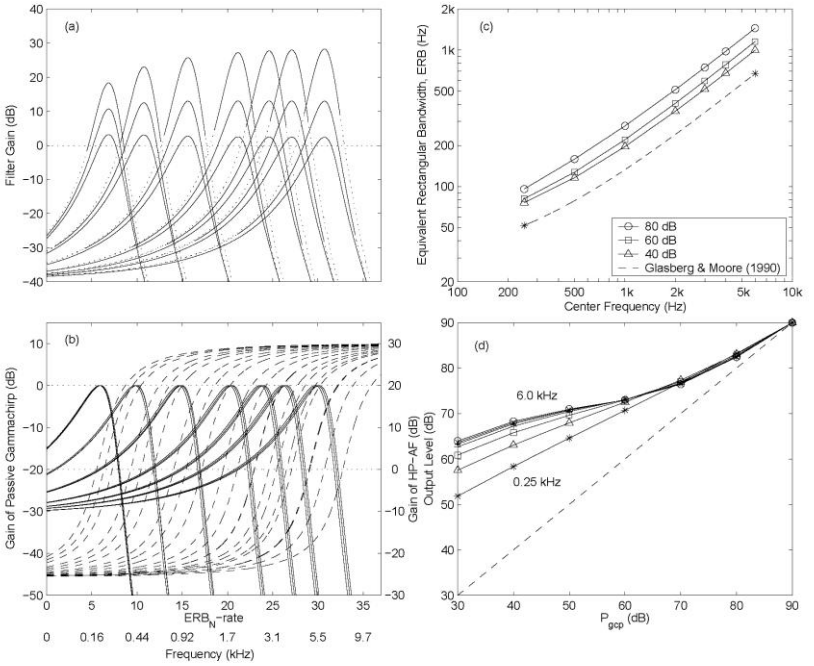


図 1・10　圧縮型 Gammachirp フィルタの諸特性

(a) フィルタ形状，(b) 受動的 gammachirp フィルタと高域通過型非対称関数，

(c) 帯域幅，(d) 入出力特性．信号周波数は 0.25, 0.5, 1k, 2k, 3k, 4k, 6 kHz **である**

一方，生理学的アプローチから提案された聴覚フィルタバンクモデルとして，基底膜上の進行波を直接模擬した Cascade/Parallel 型のフィルタバンク（Q 値一定のモデル [13]や可変 Q 値

モデル[14,15])，聴覚末梢系を伝送線路系とみなしたアナログ電子回路モデル[16]や wave digital filter によるディジタル回路モデル[17]もある．最近では，基底膜応答を線形経路／非線形経路に分けて模擬する DRNL（dual resonance nonlinear）フィルタ[18]も提案されており，順向性マスキングデータやパルセーション閾値データといった心理物理データも説明可能なフィルタバンク[19]に発展している（**注**：古典的なモデルとして Goldstein の MBPNF モデル[20] がこの構成に近い）．

ここで取り上げた聴覚モデルについては，**表 1･1** に示す URL よりフリーソフトウェアを入手可能である．特に Auditory Toolbox や AIM, DSAM パッケージには図 1･9 に示すような処理ブロックごとに必要なモデル（あるいはモジュール）を選択して一連の蝸牛フィルタリング処理をシミュレートできる環境が整っている．本節で取り上げた聴覚フィルタモデルや聴覚末梢モデルのほとんどはこれらのサイトから入手可能である．

表 1･1　フリーで入手可能な聴覚モデルのソフトウェア

Auditory Toolbox	http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/
AIM	http://www.pdn.cam.ac.uk/groups/cnbh/research/aim.html
DSAM	http://www.pdn.cam.ac.uk/groups/dsam/
Gammatone-wavelet	http://www.jaist.ac.jp/~unoki/Gammatone/index.html
HUTear	http://www.acoustics.hut.fi/software/HUTear/
LUTEar	http://www.physiology.wisc.edu/phys735/lutear/manual.html
Auditory model for ASR[24)]	http://www2.pd.istc.cnr.it/pages/asr.html

1-6-3　聴覚モデルを利用した音声信号処理・認識

これまでに音声信号処理･認識では，音声知覚特性や聴覚特性を加味した様々な特徴量（例えばメル周波数ケプストラム係数（MFCC））や分析手法（例えば，知覚的線形予測法（PLP）や変調スペクトル分析法など）が利用されてきた（1-10 節参照）．いずれも頑健な音信号処理を実現するために利用されてきたものである．これに対し，私たちの「耳」のフロントエンドである聴覚末梢系を，機械におけるフロントエンドとして利用することで，ヒトのスペクトル分析機能に匹敵する信号処理を実現することができる．例えば，聴覚モデルを低ビットレートの音声符号化[21]に利用するものや音声エンハンスメント[22, 23]に利用するものが報告されている．また，音声認識のフロントエンドとして利用することで耐雑音性が向上するといった報告[21,24,25]もある．現段階では，音声符号化や音声強調処理などで非常に高い効果が得られているが，音声認識に限ってはバックエンドの処理が人間の認識メカニズムを模擬しておらず，バックエンドに適した特徴を利用しない限り，高い効果を期待できない．そのため，聴覚末梢系をフロントエンドとした場合，これに適したバックエンド処理の実現を検討する必要がある．

現在までのところ，線形で受動的な聴覚フィルタバンクから，非線形で能動的な聴覚フィルタバンクの実現へと移行しつつある．聴覚モデルをフロントエンドとして活用できる音声符号化や補聴システムなどでは，聴覚の非線形性・能動性をよりヒトの特性に近づけることができるため，これらへのモデルの寄与は今後更に増大するものと予想される．

■参考文献

1) Fletcher, H., “Auditory patterns,” Rev. Mod. Phys., vol.12, pp.47-61, 1940.
2) Zwicker E. and Terhardt E., “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency,” J. Acoust. Soc. Am., vol.68, no.5, pp.1523-1525, 1980.
3) Patterson R. D. and Nimmo-Smith, I., “Off-frequency listening and auditory-filter asymmetry,” J.Acoust. Soc. Am., vol.67, no.1, pp.229-245, 1980.
4) Patterson, R. D. and Moore, B. C. J., “Auditory filters and excitation patterns as representations of frequency resolution,” Frequency selectivity in Hearing, edited by Moore, B. C. J., Academic Press, 1986.
5) Glasberg, B. and Moore, B. C. J.: “Derivation of auditory filter shapes from notched-noise data,” Hear. Res. Vol. 47, pp. 103-138 (1990).
6) Moore, B. C. J., An Introduction to the Psychology of Hearing, 5th ed. Academic, London, 2003.
7) Patterson, R.D., Holdsworth, J., Nimmo-Smith, I., and Rice, P., “SVOS Final Report: The Auditory Filterbank,” APU Report 2341, 1987.
8) Irino, T. and Patterson, R.D., “A time-domain, level-dependent auditory filter: The gammachirp,” J. Acoust. Soc. Am., vol.101, no.1, pp.412-419, 1997.
9) Irino, T. and Patterson, R.D., “A compressive gammachirp auditory filter for both physiological and psychophysical data,” J. Acoust. Soc. Am., vol.109, no.5, pp.2008-2022, 2001.
10) Unoki, M. Irino, T., Glasberg, B., Moore, B.C.J., and Patterson, R.D., “Comparison of the roex and gammachirp filters as representations of the auditory filter,” J. Acoust. Soc. Am., vol.120, no.3, pp.1474-1492, 2006.
11) Irino, T. and Patterson, R.D., “A dynamic compressive gammachirp auditory filterbank,” IEEE Trans. Audio, Speech, and Language Processing, vol.14, no.6, pp.2222-2232, 2006.
12) Unoki, M., Miyauchi, R., and Tan, C.-T., “Estimates of tuning of auditory filter using simultaneous and forward notched-noise masking,” Hearing – from sensory processing to Perception edited by Kollmeier, B. et al. , Springer Verlag, Heidelberg, pp.19-26, 2007.
13) Seneff, S., “A joint synchrony/mean-rate model of auditory speech processing,” J. Phonet., vol.16, pp.55-76, 1988.
14) 平原達也, “適応 Q 型非線型蝸牛フィルタ,” 日本音響学会誌, vol.47, no.5, pp.327-335, 1991.
15) Kates, J.M. “A time-domain digital cochlear model,” IEEE Trans. Signal Process., vol.39, no.12, pp.2573-2592, 1991.
16) Schroeder, M.R., “Models of Hearing,” Proc. IEEE, vol.63, no.9, pp.1332-1350, 1975.
17) Giguère C. and Woodland, P.C., “A computational model of the auditory periphery for speech and hearing research. I. Ascending path,” J. Acoust. Soc. Am., vol.95, pp.331-342, 1994.
18) Meddis, R., O’Mard, L.P., and Lopez-Poveda, E.A., “A computational algorithm for computing nonlinear auditory frequency selectivity,” J. Acoust. Soc. Am., vol.109, no.6, pp.2852-2861, 2001.
19) Lopez-Poveda, E.A. and Meddis, R., “A human nonlinear cochlear filterbank,” J. Acoust. Soc. Am., vol.110, no.6, pp.3107-3118, 2001.
20) Goldstein, J.L., “Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering,” Hear. Res., vol.49, pp.39-60, 1990.
21) Ghitza, O., “Auditory models and human performance in tasks related to speech coding and speech recognition,” IEEE Trans. Speech and Audio Processing, vol.2, no.1, Part II, pp.115-132, 1994.
22) Cooke, M.P., “Modelling Auditory Processing and Organization,” Cambridge University Press, 1993.
23) Irino, T. Patterson, R.D., and Kawahara, H., “Speech segregation using an auditory vocoder with event-synchronous enhancement,” IEEE Trans., Audio, Speech, and Language Processing, vol.14, No.6, pp.2212-2221, 2006.
24) Cosi, P., “Auditory modeling and neural networks,” in A course on speech processing, recognition, and airtificial neural networks, Springer Verlag, Lecture notes in computer science , 1998.
25) Cooke, M., Green, P., Josifovski, L., and Vizinho, A., “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Commun., vol.34, pp.267-285, 2001.

■2群 － 7編 － 1章

1-7 補　聴

（執筆者：坂本修一，鈴木陽一）［2009年5月 受領］

高齢化の更なる進展を考えたとき，自然で聞きやすい聴覚支援システムの実現は重要な課題である．補聴器は，聴覚支援システムとして最も一般的なシステムであり，これまで様々な方式が提案，市販されてきた．ここでは，現在主流となっているラウドネス補償処理（レベル圧縮処理，非線型増幅などともいわれる）に至るまでのこれまでの補聴器の流れと，近年注目されている補聴処理技術について概観する．

なお，特に日本では，話速変換を実装したラジオが市販されたり，携帯電話にラウドネス補償に類似した音声信号処理が実装されたりなど，補聴器以外の分野における聴覚支援システムの開発も進んでいる．したがって，補聴システムやアルゴリズム開発を進めるうえでは，高齢者一般を対象とした聴覚支援システムという観点でも考えることが重要である．

1-7-1 線形増幅処理からラウドネス補償処理へ

補聴器として最も古く単純なものはラッパ型の集音管を用いた集音管補聴器である．17世紀にはこのようなシステムが既に紹介されている．その後，カーボンマイクロホンを用いた補聴器（20世紀初頭），真空管・トランジスタを用いたアナログ補聴器（20世紀中頃）を経て，1995年に世界で初めて完全ディジタル補聴器の市販が開始されて以降，現在ではディジタル信号処理技術を駆使したディジタル補聴器が主流となっている．

アナログ補聴器までの補聴処理は，入力信号を単純に増幅する線形増幅処理が主であった．線型増幅補聴器では，平均的な入力レベルの音がちょうど良いラウドネスになるように調整するのが標準的方略となる．そのための方法として，ハーフゲイン法 [1]，POGO 法 [2]，NAL 法 [3]などが提案されてきた．これらの方法による増幅利得は，おおむね，健聴者と当該難聴者の聴覚域値（最小可聴値）のレベル差の半分程度の値となる．

このような値に設定する理由は，難聴者の大部分を占める感音性難聴者のほとんどが，ラウドネス補充現象（リクルートメント）を示すことにある．すなわち，感音性難聴者では，聴覚域値は健聴者に比べて高い一方，うるさくて聞きたくないレベル（不快域値）は健聴者とそれほど変わらない．そのため，音のレベルが聴覚域値を超えると，音の感覚的な大きさ（ラウドネス）が急激に上昇する（補充される）．言い換えると，感音性難聴者の聴野は健聴者に比べてかなり狭いものとなる（この現象は，内示の外有毛細胞の機能不全に起因する）．そのため，線形増幅処理ではすべての入力信号を聴野に適切に入れることはできず，平均的な音声レベルの入力をちょうど聞きやすい出力レベルにするには，聴覚域値のレベル差よりもかなり小さな利得とする必要があることとなる．その結果，線型増幅処理では，平均的なレベルの音はちょうどよい増幅が行われても，小さい音はほとんど聞こえず，逆に大きい入力音はうるさくなりすぎるという問題が生ずる．

これを解決する補聴処理がラウドネス補償処理 [4]である．**図 1･11** にラウドネス補償処理の概念図を示す．ラウドネス補償処理では，健聴者と難聴者でラウドネスが一致するように難聴者の聴野に入力信号を増幅する．具体的には，あらかじめ難聴者に対し，入力信号のレベル変化に対するラウドネスの変化（ラウドネス関数）を測定し，健聴者のラウドネス関数と

対応づけることにより，増幅率を決定する．多くの場合，入力信号を複数の帯域に分割し，帯域ごとに入力レベルに応じて増幅率を決定する．この方法は，Bell Lab.の J. B. Allen により開発され，リサウンド社のアナログ補聴器に初めて実装されたが，現在ではほとんどのディジタル補聴器が備える標準的補聴処理となっている．また，この方法の普及は，補聴を単なる増幅ではなく，難聴の補償ととらえる考え方が有効であることを示しているといえよう．

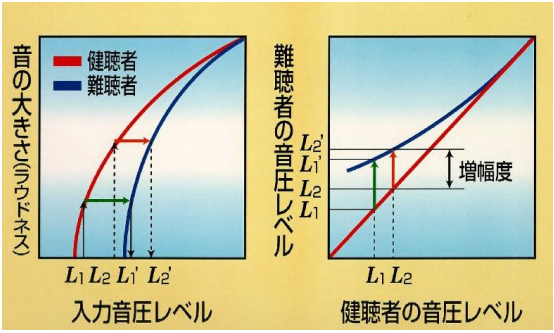


図 1・11　ラウドネス補償処理の概念図

1-7-2 選択的両耳聴アルゴリズム

基本的な補聴処理方式以外にも，様々な処理が提案，実装されている．この中でも最も一般的に用いられている補聴処理は，音声強調処理や雑音抑圧処理といった，騒音下での音声聴き取りを助ける処理である．これは，広く普及したラウドネス補償処理により，静かな場面での音声聴き取りに関してはある程度満足のいく性能が達成されている一方で，騒音下や複数の人が存在する中では，所望の音声が「聞こえはするが，内容は分からない」という状況がしばしば聞かれることにも関連する．

このような処理を実現する手法としては，スペクトル減算法，適応フィルタを用いた方法など，特に 1 入力 1 出力を考慮した手法と，独立成分分析処理，マイクロホンアレイを用いた手法など，多入力多出力を想定したシステムに大別することが可能である．両耳に着用した補聴器間の通信が現実のものとなった現在では，多入力多出力のシステムの中でも，2 入力 2 出力の補聴処理方式が極めて有望であると思われる．これは，人間が本来持っている両耳での空間情報の取得能力（選択的両耳聴）の利用が見込めるという利点も有している．

2 入力 2 出力の補聴処理方式を考えるにあたり，聴覚支援システムの両耳着用を念頭にした，両耳密結合化音声信号処理技術は重要である．両耳 2 チャネル独立ではなく，信号を相互にやりとりしながら処理を行う信号処理には多くの可能性がある [5,6]．例えば，所望の信号「のみ」を抽出して提示するのではなく，目的音を“周囲の音をある程度残しながら”聞きやすいかたちで提示する技術も重要であろう．これを更に進め，「周囲の音もある程度聞くことができる状態で，所望の音源情報が十分に聴取可能となる条件」を明らかにし，それに根ざした聴覚支援技術が開発できれば，高齢者や難聴者にとって大きな福音となるだろう．

■参考文献

1) S.F. Lybarger, “U.S. Patent Application,” SN 543, 278, 1944.
2) G.A. McCandless and P.E. Lyregaard, “Prescription of gain/output (POGO) for hearing aids”, Hear. Instrum., vol.34, no.1, pp.16-21, 1983.
3) D. Byrne and W. Tonnison, “Selecting the gain in hearing aids for persons with sensorineural hearing impairments”, Scand. Audiol., vol.5, pp.51-59, 1976.
4) 浅野太，鈴木陽一，曽根敏夫，林哲也，佐竹充章，大山健二，小林俊光，高坂知節，“ラウドネス補償特性を有するディジタル補聴器の一構成法，”日本音響学会誌，vol.47, no.6, pp.373-379, 1991.
5) H. Nakashima, Y. Chisaki, T. Usagawa and M. Ebata, “Frequency domain binaural model based on interaural phase and level differences,” Acoust. Sci. & Tech., vol.24, no.4, pp.172-178, 2003.
6) J. Li, S. Sakamoto, S. Hongo, M. Akagi and Y. Suzuki, “Advancement of two-stage binaural speech enhancement (TS-BASE) for high quality speech communication,” Proc. WESPAC X, 184, 2009.

■2群 － 7編 － 1章

1-8 音声言語の知覚

（執筆者：天野成昭）［2009年4月 受領］

1-8-1 音韻知覚

(1) ボトムアップ情報

音韻知覚は音声波の音響物理的特徴を利用して行われる．この音響物理的特徴を，後述のトップダウン情報に対応させてボトムアップ情報と呼ぶ．音響物理的特徴は多重性，分散性，変動性などの性質を持つ．

(a) 多重性

各音韻の知覚のキューとなる音響物理的特徴は多重に存在し，しかも各特徴の重要度は同等ではなく，それぞれ異なっている．例えば/p/の音響物理的特徴には，先行母音末尾のホルマント遷移パターン，無音区間，短いバースト性のノイズ，ノイズの開始から後続母音の開始までの時間長（voice onset time: VOT），後続母音開始部のホルマント遷移パターンなどが挙げられる．これらのうち，先行母音末尾や後続母音開始部のホルマント遷移パターンは，重要度が高い音響物理的特徴ではない．なぜならば，語頭・文頭など先行母音が存在しない場合や，無声化によって後続母音が存在しない場合でも，/p/の知覚が可能だからである．音響物理的特徴の多重性のメリットの一つは，いくつかの音響物理的特徴が背景雑音などによって失われた場合でも，他の音響物理的特徴の組合せによって，本来の音韻が知覚可能となる点である．

(b) 分散性

各音韻の知覚のキューとなる音響物理的特徴は，時間上に連続的に分散して存在する（**図1･12**）．このため音声波形上で，各音韻の開始時点や終了時点を定めることは一般に簡単ではない．連続音声中から，/a/や/i/に相当する部分の波形を切り出し，それらを接続して/ai/を合成した場合，その認識率や明瞭性や自然性は，連続音声の/ai/に比べて劣化するのが普通である．これは連続的に分散表現された音響物理的特徴の一部が波形の切り出しによって失われ，かつ接続部分の前後において分散表現に不整合が生ずるためだと考えられる．

各音韻の音響物理的特徴が時間上に連続的に分散して存在するという性質にもかかわらず，連続音声を聴取すると，音韻が徐々に変化するような知覚は生じず，離散的な音韻の列が知覚される．音声知覚システムは，連続的に分散して存在する音響物理的特徴から離散的な音韻のシンボル列を取り出しているといえる．

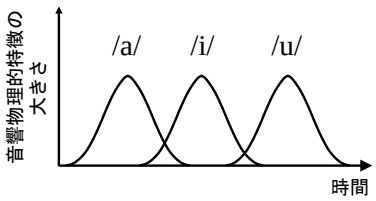


図 1･12 音響物理的特徴の分散性の概念図

(c)　変動性

各音韻の知覚のキューとなる音響物理的特徴は，その前後に存在する音韻の種類や発声速度などの文脈によって変動する．例えば，破裂子音の音響物理的特徴の一つである VOT は，遅い発声速度では長く，早い発声速度では短くなる [1]．各音韻の音響物理的特徴は，調音器官の形状と大きさ，及びその動作パターンに依存して変動する．調音器官の形状と大きさには個人差や男女差や年齢差が存在し，またその動作パターンにも個人差が存在するので，その差に対応して各音韻の音響物理的特徴にも差が生ずる．音声知覚システムは，このような各種の変動を考慮に入れて，頑健な音韻の知覚を達成していると考えられる．

(2)　トップダウン情報

音韻知覚ではボトムアップ情報に加えてトップダウン情報も利用される．トップダウン情報とは経験や学習に基づいて脳内に記憶された情報であり，音韻配列規則情報，心的辞書情報，構文情報，意味情報などが例として挙げられる．

心的辞書に存在する単語中の音韻は，心的辞書に存在しない非単語中の音韻に比べて検出されるまでの時間が短く [2]，また知覚されやすい．例えば，語頭破裂子音の VOT を系統的に変化させて合成した「単語/dash/－非単語/tash/」の刺激連続体，及び「非単語/dask/－単語/task/」の刺激連続体を聴取して/d/と/t/のカテゴリー境界を求めると，そのカテゴリー境界は通常の/d/と/t/のカテゴリー境界よりも，非単語側（すなわち/tash/や/dask/側）になる [3]．これは/d/と/t/の間のあいまいな子音を持つ刺激が，心的辞書情報によって単語（すなわち/dash/や/task/）として知覚されやすくなったことを示している．

各種のトップダウン情報の利用はバイアスとして働き，聞き間違いを引き起こすこともある．しかし一般的には，ボトムアップ情報だけでは音韻が知覚不能である劣悪な音環境においても，不足する情報を補って音韻を知覚可能とするというメリットの方が大きい．つまりトップダウン情報の利用によって音韻知覚の頑健性が増す．

1-8-2 音声単語認知

音声波から抽出された音韻列と心的辞書中の単語との照合がなされ，音声単語が認知される．この認知過程には，構文情報や意味情報などが文脈として影響を及ぼすほか，単語自体が持つ性質，例えば単語のなじみの程度（単語親密度）[4]なども影響を及ぼす．以下に音声単語認知過程の代表的モデルであるコホート・モデルとトレース・モデルを示す．

(1)　コホート・モデル

コホート・モデル（cohort model）[5]の特徴は，音声単語認知過程において活性化される単語候補をコホートと呼ばれる群に限定した点，及びそのコホートにおける音韻を単位とした逐次的照合処理を仮定した点である．入力される音声が /benifit/（benefit）であったとき，最初に語頭部分が /b/ であるすべての単語が単語候補群（コホート）として活性化される（**図 1･13** a）．次に，入力音声と各単語候補との逐次的照合処理が，時間軸に沿って行われる．すなわち，音韻が一つ入力されるごとに，当該位置においてその音韻と一致しない音韻を持つ単語候補が次々と削除され，コホートが絞り込まれてゆく（図 1･13 b, c）．

この絞り込みには意味などの文脈情報も利用され，意味的に不適切な単語候補はコホートから削除される．このような絞り込みの処理の結果，最終的にコホート中に単語候補が一つ

だけ残る（**図 1･13** d）．この時点をユニークネス・ポイント（uniqueness point: UP）と呼び，この時点において単語認知が起きる．つまり，コホート・モデルで単語認知が起きるためには，必ずしも単語の末尾まで照合を行う必要はない．

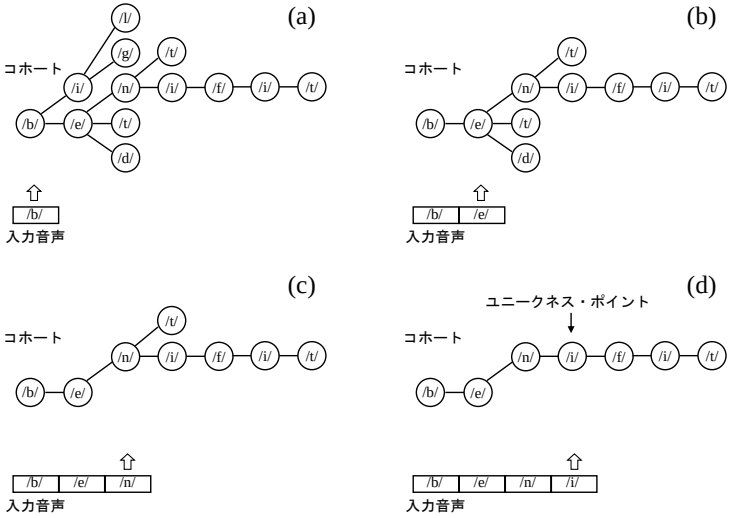


図 1･13　コホート・モデル

(2)　トレース・モデル

トレース・モデル（TRACE model）[6]の特徴は，ニューラルユニットによる並列分散処理を仮定した点である．トレース・モデルは音響特徴処理レベル，音韻処理レベル，単語処理レベルの 3 層で構成される（**図 1･14**）．音響特徴処理レベルには母音性，子音性，有声性，破裂性，集約性，高音調性等の音響特徴に対応した音響特徴ユニットが存在する．音韻処理レベルには各音韻に対応した音韻ユニットが存在し，単語処理レベルには各単語候補に対応した単語ユニットが存在する．各ユニットは活性度，静止状態の値，及び減衰時間の値を持つ．ユニット間の結合は重みを持ち，双方向性である．

上下関係にあるレベル間においては，ユニット間の結合が促進性である．各ユニットはこの促進性結合を通して他のレベルのユニットから入力を受け活性化する．例えば，単語処理レベルに存在する単語ユニットの /big/ は，音韻処理レベルに存在する音韻ユニットの /b/, /i/, /g/ から促進的入力を受けて活性化する．結合は双方向性であるので，これとは逆方向の活性化も生ずる．すなわち，音韻ユニットの /b/, /i/, /g/ は，単語ユニットの /big/ から促進的入力を受けて活性化する．一方，各レベル内においてはユニット間の結合が抑制性である．各ユニットは，この抑制性結合を通して，同レベルに存在する他のユニットの活性化を妨げる．例えば，単語処理レベルに存在する単語ユニットの /big/ と /it/ は，抑制性結合を通して相互に相手の活性度を抑制する．このようなレベル間及びレベル内の相互作用を続けるこ

とによって，最終的に各レベルにおいて，入力音声に最も適したユニットだけが高い活性度となり，その他のユニットはすべて低い活性度になる．単語認知は，ある単語ユニットの活性度が他のすべての単語ユニットよりも大きくなり，その活性度から計算した確率が特定の値を越えたときに起きる．

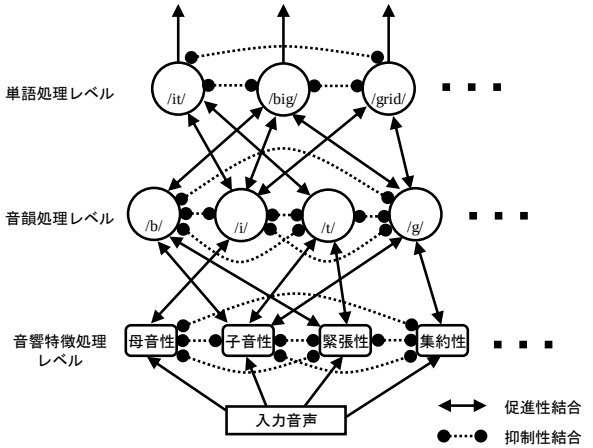


図 1・14　トレース・モデル

トレース・モデルの各ユニットが入力を受け取る時間範囲には幅があり，上位のレベルほどその幅は大きい．これによって，トレース・モデルは時間的に分散して存在する特徴をとらえることができ，時間軸上で先行する文脈ばかりでなく後続する文脈も扱えるようになっている．またトレース・モデルには，各時刻において，各ユニットのコピーが時間的に重複して存在する．これによって，入力がどの時刻に存在しても，それに応じたユニットが活性化するようになっている．

■参考文献

1) K. Nagao and K. de Jong, “Perceptual rate normalization in naturally produced rate-varied speech,” The Journal of the Acoustical Society of America, vol.121, pp.2882-2898, 2007.
2) S.Amano, “Effects of lexicon and coarticulation on phoneme perception,” The Journal of the Acoustical Society of Japan (E), vol.14, pp.91-97, 1993.
3) W.F. Ganong III, “Phonetic categorization in auditory word perception,” Journal of Experimental Psychology: Human Perception and Performance, vol.6, pp.110-125, 1980.
4) 天野成昭，近藤公久，“日本語の語彙特性，”第 1 巻，三省堂，東京，1999.
5) W. Marslen-Wilson and A. Welsh, “Processing interactions and lexical access during word recognition in continuous speech,” Cognitive Psychology, vol.10, pp.29-63, 1978.
6) J.L. McClelland and J.L. Elman, “The TRACE model of speech perception,” Cognitive Psychology, vol.18, pp.1-86, 1986.

■2 群 － 7 編 － 1 章

1-9 音声中の非言語情報の知覚

（執筆者：北村達也）［2009 年 5 月 受領］

音声は，文字に書き起こすことができる情報（狭義の言語情報）のみならず，性別，年齢，個人性，感情，意図，体調など様々な非言語情報を伝達する．これらの情報は，容貌や表情によって伝達される情報と似通っていることから auditory face と呼ばれることがある．我々の音声言語コミュニケーションではこれらの非言語情報も重要な役割を担っている．

音声は，声帯や声道内にて生じた音源が声道を通過し，空気中に放射されることによって生成される．したがって，音声中の非言語情報の知覚においては，声帯音源や声道の音響特性における話者間の差異や話者内の変化に由来する音響的特徴が重要な手がかりになる．

1-9-1 性別の知覚

成人の音声において性別の知覚に大きく寄与するのは，音声の基本周波数の性差である．日常会話の平均的な基本周波数は，成人男性で 120 Hz，成人女性で 240 Hz であり，成人男性は成人女性より話声の声域が約 1 オクターブ低い．男性の場合，思春期に喉頭の軟骨が発達するのに伴い声帯の長さと質量が増加し，声帯振動の周波数低下を引き起こす．これが変声（声変わり）である．女性においても変声は生ずるものの，基本周波数の低下は小さい [1]．

声道長の性差に由来する声質の違いも性別の知覚に寄与する．一般に，男性の声道は女性よりも長い．声道長に性差が現れるのも変声期である．声道長が長くなるほど音声スペクトルのピークは低周波数側にシフトし，これが声質の違いを生み出す．

1-9-2 年齢の知覚

変声前の子供は成人と比較して声帯長と声道長が共に短いため，話声の声域が高く，音声スペクトルのピークが高周波数側にシフトしている．また，調音運動機能が未発達なことによるたどたどしさも子供の音声を特徴づける．

声帯は，個人差はあるものの加齢とともに質的，量的な変化が起こるため，声帯振動にも影響が生ずる．そのため，一般的に，男性では変声期以後 50 代後半まで話声の基本周波数がほぼ変化しないが，60 代から上昇し，女性では年齢とともに基本周波数が低下する [2]．したがって，特に女性では基本周波数が年齢の知覚に寄与していると考えられる．

高齢者の音声では基本周波数以外にも独特の特徴を有している．その特徴を表す形容語として，「しゃがれ」，「不明瞭さ」，「発話の遅さ」の 3 因子が報告されている [3]．

1-9-3 個人性の知覚

音声の個人性は，長時間平均音声スペクトル，持続音声のスペクトル，平均基本周波数などの時間的に静的な成分に現れるものと，基本周波数や音声スペクトルの時間的に動的な成分に現れるものに大別できる．前者は主に発話器官の生得的な特徴に由来し，後者は主に学習によって後天的に獲得されたものである．個人性の知覚にはこれらの音響的特徴が寄与している．ただし，個人性が顕著に現れる特徴は話者ごとに異なり，また，個人性知覚の際に重視される特徴も個人ごとに異なると考えられる．

個人性知覚のしやすさには音韻による違いがあり，母音や鼻音ではこれら以外の音韻よりも容易であることが知られている．これらの音韻では声帯音源と声道（鼻音の場合は鼻腔を含む）の音響特性の個人差が音声に強く反映されるためである．

時間的に静的な成分のうち，長時間平均音声スペクトルと持続音声のスペクトルにおける個人差は，声道長と声道形状の個人差に由来し，個人性知覚に寄与する．一例を挙げると，声道下部の形状の個人差は，母音の個人性知覚への寄与が知られている音声スペクトルの高域成分（約 2.5 kHz 以上）に個人差をもたらす[4]．また，平均基本周波数における個人差は，主に発声器官の個人差による．このほか，「太い声」，「かすれた声」などの特徴も発声器官の個人差に由来し，個人性知覚に寄与する．

継続時間が等しい持続母音と連続音声の話者識別のしやすさを比較してみれば分かるように，時間的に動的な成分も重要な知覚要因である．訛りや話し方の癖もこの分類に含まれる．

個人性と密接に関連する声質に関する形容語としては，六つの表現語対，「高い声－低い声」，「かすれた声－澄んだ声」，「落ち着きのある声－ない声」，「迫力のある声－弱々しい声」，「太い声－細い声」，「張りのある声－ない声」，及び一つの表現語「鼻声」が報告されている[5]．

脳機能障害の一種に，話者の識別ができない phonagnosia と呼ばれる症例がある[6]．この患者は音声知覚には問題がないものの，話者の識別ができない．このような患者の存在は，脳内の独立したメカニズムによって個人性が処理されていることを明示している．

1-9-4 感情の知覚

上述の非言語情報が話者間の差異に関する特徴であったのに対して，感情は話者内の変化に関する特徴である．感情の知覚には，基本周波数，話速，ラウドネス，声質が複合的に寄与する[7]．これらの特徴の変化は，話者による意識的，無意識的な声帯音源と声道の音響特性の操作によってもたらされる．

例えば，怒りの感情を表した音声では，平静時の音声と比較して，話速が速くなり，ラウドネス及びその変化も増大する．更に，基本周波数も上昇し，その変化も増大する．一方，悲しみの感情を表した音声では，怒りの音声とは逆の特徴が見られる．これらの特徴が感情知覚の手がかりとなる．

近年では，聞き手の文化的背景が感情知覚に及ぼす影響が認識され，様々な文化間の調査が行われるようになっている[7]．

■参考文献

1) 日本音声言語医学会，“新編声の検査法，”医歯薬出版，2009.
2) 粕谷英樹，森大毅，木戸博，“年齢による話声の基本周波数の変化，” 音講論，pp.281−282, Sep. 2006.
3) 宮崎健，水町光徳，二矢田勝行，“高齢者音声の聴覚的特徴を形容する語の抽出に関する検討，”信学技報（SP），vol.108, no.47–52, Oct. 2008.
4) T. Kitamura, K. Honda and H. Takemoto, “Individual variation of the hypopharyngeal cavities and its acoustic effects,” Acoust. Sci. & Tech., vol.26, pp.16–26, Jan. 2005.
5) 木戸博，粕谷英樹，“通常発話の声質に関連した日常表現語: 聴取評価による抽出，”音響論，vol.57, pp.337−334, May 2001.
6) L. Garrido, F. Eisner, C. McGettigan, L. Stewart, D. Sauter, J.R. Hanley, S.R. Schweinberger, J.D. Warren and B.Duchaine, “Developmental phonagnosia: A selective deficit of vocal identity recognition,” Neuropsychologia, vol.47, no.123–131, Jan. 2009.

7) D. Erickson, “Expressive speech: Production, perception and applicaiton to speech synthesis,” Acoust. Sci. & Tech., vol.26, pp.317–325, April 2005.

■2群 － 7編 － 1章

1-10 音声分析法

（執筆者：河原英紀）［2009年8月 受領］

音声認識や合成の様々な手法が提案されている．音声分析法は，それぞれの方法に適したかたちで表現された音声情報を提供する．この観点に立つなら，アナログ信号のディジタル化の新しい流れ [1]が，広義の音声分析法に含まれる日も遠くはない．それらの最近の動向を意識しつつ，ここでは基礎的な考え方から順を追って，音声分析に用いられる手法の紹介を進めることとする．

1-10-1 フーリエ変換に基づく表現

声門などで生じた音源は，声道の共鳴と唇からの放射の影響を受けることで音声となる．このようなモデルを，線形時不変システムとして近似する．すると，観測される音声のスペクトル S (ω) は，次式に示すように，音源波形のスペクトル G(ω)，声道の伝達特性 H(ω)，唇からの放射特性 R(ω) の積として表される．

$$S(\omega) = R(\omega)H(\omega)G(\omega) \tag{1・9}$$

ここでωは，角周波数を表す．この両辺の対数を求めることで，音声の対数スペクトルを，それぞれの要素の対数スペクトルの和に分解することができる．このように，フーリエ変換と対数による処理は，一見すると複雑な音声波形を，見通しの良いものにする．

実際の音声では，声道形状は時間とともに変化する．そのため，応用の際には，音声波形$x(t)$を窓関数$w(t)$により局所的に観測した短時間フーリエ変換$S(\omega, t)$が用いられる．以下では特に断らない限り，音声のスペクトルは短時間フーリエ変換を指すこととする．

$$S(\omega, t) = \int_{t-T/2}^{t+T/2} x(\tau)w(\tau - t)e^{-j\omega\tau}d\tau \tag{1・10}$$

ここで t は，観測のための窓関数の位置，T は，窓関数の長さを表す．なお，音声知覚への位相の影響は少ないため，絶対値にのみ依存するパワースペクトルが用いられることが多い．このようにして求められる時間周波数表現$|S(\omega, t)|^2$は，スペクトログラムと呼ばれる．

スペクトログラムは，現在でも音声の可視化の手段として広く用いられている．窓長を短くすることで求められる，ホルマントや男性の音声の周期性などの特徴を観測しやすい広帯域スペクトログラムと，窓長を短くすることで求められる，音声の調波構造を観測しやすい狭帯域スペクトログラムとが，用途に応じて使い分けられる．

(1) ケプストラム，複素ケプストラム

母音や鼻子音のような有声音では，声門は周期的な開閉運動を繰り返している．このような繰り返しがあると，音声のスペクトルには$\omega_0 = 2\pi f_0$を周期とする変動が乗算のかたちで加えられる．ここで，f_0 は，基本周波数（F0）を表す．音声スペクトルの対数のフーリエ変換を求めると，フーリエ変換の線形性により，それぞれの成分の対数変換のフーリエ変換の和として表される．ケプストラム（Cepstrum）分析は，この性質を利用している．

複素ケプストラム$c(q)$とケプストラム$C(q)$は，音声の対数スペクトルから次式により求め

られる[2].

$$c(q) = \int \log(S(\omega))e^{j\omega q}d\omega \quad (1\cdot 11)$$

$$C(q) = \int \log|S(\omega)|e^{j\omega q}d\omega \quad (1\cdot 12)$$

ここで用いられる変数 q は，ケフレンシ（quefrency）と呼ばれ，時間間隔に対応している．音声の周期性に起因する成分は，高いケフレンシにピークを有し，声道伝達特性や音源波形の形状と放射特性に起因する成分は，低いケフレンシ領域に集中している．この性質を利用して，ケプストラムは，F0 抽出や音声合成におけるスペクトル包絡の抽出に用いられている．後述の統計的な観点と聴覚における周波数分析の性質を取り入れたメルケプストラム[3]は，音声合成のためのパラメータとして広く用いられている．

(2) フィルタバンク

ある ω_t での短時間フーリエ変換の値 $S(\omega_t, t)$ は，インパルス応答が $w(t)\exp(j\omega_t t)$ であるような帯域通過フィルタの出力と解釈することができる．このようなフィルタを，ある規則に基づいて配置したものをフィルタバンクと呼ぶ．短時間フーリエ変換は，等帯域幅のフィルタを，中心周波数が周波数軸上で等間隔になるように配置したフィルタバンクによる分析に相当する．中心周波数に対する帯域幅の比が一定となる帯域通過フィルタを対数周波数軸上で等間隔に配置したフィルタバンクは，ウェーブレット変換に相当する．内耳における音の分析は，短時間フーリエ変換よりもウェーブレット変換により近い．

(3) MFCC

音声認識では，このフィルタバンクによる処理とケプストラム分析を組み合わせて求められる MFCC（Mel Frequency Cepstral Coefficient）[4]が広く用いられている．MFCC に用いられるフィルタバンクでは，聴覚における周波数分析を模した非直線周波数軸（Mel frequency）上に，三角形の特性を持つ帯域フィルタが配置されている．MFCC は，それらのフィルタ出力のパワーの対数を，離散コサイン変換することにより求められる．なお，実際の認識系では，隣接する分析フレームにおける MFCC の差分として定義されるΔMFCC や，更に，その差分として定義されるΔΔMFCC が，MFCC と併せて用いられることがある．

(4) 離散信号と FFT

フーリエ変換に基づく処理は，高速フーリエ変換（Fast Fourier Transform: FFT）[5]の発明により，急速に普及した．離散フーリエ変換をそのまま実装した場合に必要となる $O(n^2)$ のオーダーの積和の計算量が，FFT を用いた場合には $O(n \log n)$ のオーダーとなる．そのため，フィルタ処理のための畳込みや相関関数の計算に FFT を用いることにより，計算量を大きく削減することができる[6]．上記の分析も，現在では FFT を用いて実装されることが多い．

短時間フーリエ変換のために，様々な窓関数が用意されている．音声認識のための分析では，メインローブが鋭い Hamming 窓が用いられることが多く，音声合成のための分析では，サイドローブが低い Blackman 窓などが用いられることが多い．矩形窓に換算した窓関数の実質的な長さでは 15 ms 程度，位置の更新周期として，認識では 10 ms，合成では 5 ms のものがよく用いられる．

1-10-2 統計的性質に基づく表現

同じ内容の言葉を発しても，観測される音声信号は，毎回，不規則に変化する．これらの不規則な変化の背景にある共通の性質を調べるためには，観測される音声信号を，ある確率過程に属する標本信号として扱うことが必要となる．以下では，一次のモーメントと二次のモーメントのみで記述できる弱定常過程を音声信号のモデルとする．信号の平均値が 0 である場合には，次式で定義される信号の自己相関関数 $r[n_1, n_2]$ より，性質が記述される．

$$r[n_1, n_2] = E\{x[n_1]x[n_2]\} \tag{1・13}$$

ここで $x[n]$は，離散化された時刻 n における信号の観測値を表す．弱定常過程の場合には，$r[n_1, n_2]$は，時間差の関数 $r[n_1 - n_2]$となる．自己相関関数は，パワースペクトル $P(\Omega)$と次の関係により結びつけられている．

$$P(\Omega) = \sum_{m=-\infty}^{\infty} r[m]e^{-j\Omega m} \tag{1・14}$$

ここでΩは，信号の標本化周期により正規化された角周波数であり，$-\pi$からπの間の値をとる．

(1)　LPC

この弱定常過程として，現在の値が過去の値の線形結合とランダムな入力によって決まる自己回帰過程を仮定し，パラメータを推定する方法が LPC（Linear Prediction Coefficient）分析[7, 8]の基礎となっている．板倉らによる導出[7]では，$x[n]$が，以下のモデルに従うことを仮定し，最尤法によりパラメータα_mを求めている．

$$\varepsilon[n] = \sum_{m=0}^{M} \alpha_m x[n-m] \tag{1・15}$$

ここで，$\varepsilon[n]$は，$x[n]$とは独立な白色雑音である．α_mの最尤推定値は，以下の連立方程式を解くことにより求められる．

$$\sum_{m=1}^{M} \alpha_m r[k-m] = -r[k] \tag{1・16}$$

このα_mにより構成されるフィルタは，次式に示すような全極型の伝達関数 $H(z)$を持つ．

$$|H(z)|^2 = \left| \frac{1}{1 + \sum_{m=1}^{M} \alpha_m z^{-m}} \right|^2 \tag{1・17}$$

LPC 分析は，全極型の伝達特性を用いて音声のパワースペクトルを近似する際にパラメータを最尤推定する方法である．

(a)　LPC ケプストラム

こうして推定された$|H(z)|^2$ の対数をフーリエ変換することにより，波形から直接求める場合と同様に，ケプストラムを求めることができる．実際には，推定されたα_mから漸化式を用いて計算する方法が与えられている．こうして計算されるケプストラムは，LPC ケプストラムと呼ばれる．なお，波形から直接計算されるケプストラムと LPC ケプストラムの値は，異なる．

(b)　PARCOR

LPC 分析により求められる予測子係数α_m は，量子化や補間によるスペクトルへの影響が

大きい．PARCOR は，前向きの線形予測を行った場合の予測残差と，後ろ向きの線形予測を行った場合の予測残差を用いて，偏相関係数（Partial Correlation）として定義される．(p+1)次の PARCOR 係数 k_{p+1}を，p 次の予測子係数の組$\{\alpha_i^{(p)}\}(i = 1, \ldots, p)$と自己相関係数から，漸化式を用いて求める効率的な計算法が与えられている．こうして求められた PARCOR 係数の値の絶対値は 1 以下となるため，α_mよりも扱いやすい．なお，PARCOR 係数は，声道を一次元の音響管の従属接続で近似したときの，接続点での反射係数に対応する．

(c) LSP（LSF）

LSP（Line Spectrum Pair）[9]あるいは LSF（Line Spectrum Frequency）も，LPC から派生したパラメータである．m 次の PARCOR 係数を用いて合成フィルタを構成する際には，終端条件に相当する k_{m+1}を，無反射を意味する 0 に設定する．この値を 1 あるいは－1 に設定した場合には，伝達特性は線スペクトルとなる．これらの線スペクトルの周波数を順に並べたものが LSP である．LSP を求めるためには代数方程式を解く必要があるが，根の性質を利用した効率の良い計算法が与えられている．LSP の値は周波数であり，性質が理解しやすい．また，量子化特性も補間特性も良く，PARCOR の 60%の情報量で同程度の品質の音声が合成できる．これらの特長により，LSP は広く用いられる表現となっている．なお，自己相関係数，LPC, PARCOR, LSP, LPC ケプストラムは，いずれにも相互に変換が可能である．これらは，同じ内容を別のかたちで表現したものであり，変換によって新たな情報が生まれるわけではない．

1-10-3 音声合成のための分析

これまでに紹介した分析法は，様々なかたちの音声合成に用いられてきた．フィルタバンクは，最初の電気的音声合成装置の Voder や分析合成装置のボコーダ（VOCODER）[10] に用いられた．LPC 及びそれから導かれる表現や，複素ケプストラムも，同様に音声の分析合成システムに用いられてきた．これらの表現は，合成法のフィルタ部分を設計する際に用いられる．合成のためには，それらに加え，フィルタを駆動するための信号が必要となる．F0 は，駆動信号を設計するための最も重要な情報である．

これらのボコーダ型の分析合成音には特有の品質劣化があり，避けられないものと考えられていた．そのため，高い品質の音声を合成する方法としては，信号を正弦波とランダムな成分に分解し，それらを組み合わせることにより合成する方法（sinusoidal model）[11, 14]が用いられてきた．しかし，信号の周期性に注目した STRAIGHT[12, 13]は，ボコーダ型を用いても高い品質の音声を合成することが可能であることを示し，現在では合成法の選択肢が増えている．以下では，F0 の抽出法と，sinusoidal model，STRAIGHT に用いられている分析法について説明する．

(1) F0 の抽出

有声音の F0 は，声門の閉止から次の閉止までの時間間隔の逆数として定義される．周期信号の基本周期の逆数として定義される F0 は，知覚される音の高さであるピッチと密接に関連しているため，F0 抽出は，ピッチ抽出と呼ばれることが多い．しかし，合成に必要なのは F0 であり，心理量であるピッチではないことに注意する必要がある．

合成に用いるための F0 を推定する方法には，非常に多くの提案がある．それらを，基本となるアイディアに基づいて整理すると，大きく以下のように分けることができる．

(a)　基本波に基づく方法

基本周期は基本波成分の周期と一致する．基本波に基づく方法では，フィルタにより選択された基本波の零交差間隔や瞬時周波数を求め，F0 を推定する．この方法の性能は，基本波の選択に大きく依存しており，様々なフィルタの設計法などが試みられている．

(b)　波形の繰り返し間隔に基づく方法

異なった時刻で切り出された波形間の相関は，時刻の差が基本周期と一致する場合に最大となる．また，その場合，波形間の差は最小となる．波形の繰返し間隔に基づく方法では，これらの性質を用いて，F0 を推定する．推定誤りを減少させるために，前処理として，LPC による逆フィルタ処理や，スペクトル包絡の逆特性を用いた白色化が行われることが多い．

(c)　スペクトルの周期的特徴に基づく方法

ケプストラムの項で触れたように，波形の繰り返しは，スペクトルに周期的な変動を乗算的に加える．ケプストラムに基づく F0 推定は，対数スペクトルをフーリエ変換することにより基本周期に対応する成分が独立したピークとして表れることを利用している．同様な効果は，周期的な変動のあるスペクトルをスペクトル包絡を用いた除算により正規化することによっても得ることができる．

(2)　正弦波モデル（Sinusoidal model）

F0 抽出とそれに伴う有声／無声判定を高い精度で行うことは困難なことが多い．正弦波モデル[11]では，フィルタと音源を分離することをやめ，音声波形 $s(t)$を次式を用いて表す．

$$s(t)=\sum_{l=1}^{L(t)} A_l(t)\cos[\theta_l(t)] \qquad (1\cdot 18)$$

ここで $A_l(t)$には，音源波形，声道伝達特性，放射特性の影響がすべて含まれており，$\theta_l(t)$には成分の瞬時周波数及び初期位相として，音源波形や伝達特性，基本周期の変化などによる位相変化がすべて含まれている．分析では，短時間フーリエ変換により求められたスペクトルの局所的なピークの選択により $A_l(t)$が求められ，そのピークにおける位相と前後のフレームの位相との接続により $\theta_l(t)$が求められる．なお，この表現では，摩擦子音のような非周期的な部分も正弦波の和として表現されるため，正弦波成分の軌跡は，不自然な増減を繰り返す．

(a)　ランダムな成分を加えたモデル

Sinusoid plus noise model[14]では，正弦波の個数を固定し，ランダムな成分を表す項を加えることで，この問題点を解消している．このモデルでは，音声波形は以下のように表される．

$$s(t)=\sum_{l=1}^{L} A_l(t)\cos[\theta_l(t)]+r(t) \qquad (1\cdot 19)$$

ここで $r(t)$は，ランダムな成分を表す．分析では，正弦波モデルと同様の手段で正弦波成分を抽出した後，元の音声波形との差としてランダムな成分が求められる．

(3)　STRAIGHT

STRAIGHT は，背景にある滑らかな時間周波数表現を標本化する操作として，信号の周期性の役割を解釈する[12]ことにより提案された方法である．ここでは，見通しの良い新しい定式化[13]に基づいて説明する．

パワースペクトル$|S(\omega, t)|^2$を，基本周期 T_0 の半分だけ離れた二つの時刻で求めて和を計算すると，窓関数と波形の相対位置に依存する成分が相殺される．この性質を利用して，次式により相対位置に依存しないパワースペクトル $P_T(\omega, t)$がまず求められる．

$$P_T(\omega,t)=\left|S(\omega,t-T_0/4)\right|^2+\left|S(\omega,t+T_0/4)\right|^2 \qquad (1\cdot 20)$$

こうして求められた $P_T(\omega, t)$を，スペクトル包絡が周波数領域でω_0の周期で標本化され平滑化されたものと解釈する．すると，スペクトル包絡の推定は，DA 変換の問題となる．ここで，DA 変換された信号と元の信号の値が標本点において一致することだけを要請する consistent sampling[15]を利用することで，以下の近似計算法によりスペクトル包絡 $P_S(\omega, t)$が求められる．

$$\begin{aligned} L(\omega) &= \log[g(\omega)*P_S(\omega,t)] \\ P_S(\omega,t) &= \exp[q_0L(\omega)+q_1(L(\omega+\omega_0)+L(\omega-\omega_0))] \end{aligned} \qquad (1\cdot 21)$$

ここで $g(\omega)$は，積分値が 1 に正規化されたω_0の幅の矩形の平滑化関数である．q_kは，窓関数と平滑化関数の相関から計算される補償用の係数である．

1-10-4 音声のディジタル化と情報表現

これらの分析は，適切な標本化周波数と量子化ビット数の下でディジタル化された音声を対象としている．また，特に断らない限り，量子化ビット数の影響を無視した離散時間信号として扱われている．情報処理能力と MEMS 技術の急速な進化により，処理をセンサチップの中あるいは，その中の AD 変換の前のアナログ部分に出すことも可能になってきている．これは，新しい流れである Compressive sampling ともなじみが良い [1]．これらにより従来とは異なった方法でディジタル化された音声を対象とする時代は目前に迫っており，その進展に応じてこれらの前提を見直す必要がある．

■参考文献

1) Y.C. Elder, T. Michaeli, “Beyond bandlimited sampling,” IEEE Signal Processing Magazine, vol.26, no.3, pp.48–68, May 2009.
2) A. Oppenheim, R. Shafer, “Homomorphic analysis of speech,” IEEE Trans. Audio and Electroacoustics, vol.16, no.2, pp.221–226, June 1968.
3) 今井聖，“音声信号処理，” 森北出版，1991.
4) S.B. Davis, P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Trans. Acoustic, Speech and Signal Processing, vol.28, no.4, pp.357–366, Aug. 1980.
5) J.W. Cooley, J.W. Tukey, “An algorithm for the machine calculation of complex Fourier series,” Math. Comput., vol.19, pp.297–301, 1965.
6) A. Oppenheim, R. Shafer, “Discrete-time signal processing,” Prentice Hall, 1989.
7) 板倉文忠, 齋藤収三, “統計的手法による音声スペクトル密度とホルマント周波数の推定，”信学論（A），vol.53-A, no.1, pp.35–42, Jan. 1970.
8) B.S. Atal, S.L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., vol.50, no.2, pp.637–655, Aug. 1971.
9) 菅村昇，板倉文忠，“線スペクトル対（LSP）音声分析合成方式による音声情報圧縮，”信学論（A），vol.64-A, no.8, pp.599–606, Aug. 1981.

10) H. Dudley, “Remaking speech,” J. Acoust. Soc. Am., vol.11, no.2, pp.169–177, Oct. 1939.
11) R.J. McAulay, T.F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” IEEE Trans. Acoust. Speech and Signal Processing, vol.ASSP-34, no.4, pp.744–754, Aug. 1986.
12) H. Kawahara, I. Masuda-Katsuse, A. de Cheveigné,“Restructuring speech representations using a pitchadaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Comm., vol.27, no.3–4, pp.187–207, April 1999.
13) H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, “Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation,” ICASSP2008, Las Vegas, pp.3933–3936, April 2008.
14) X. Serra, J. Smith, “Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition,” Computer Music Journal, vol.14, no.4, pp.12–24, Winter 1990.
15) M. Unser, “Sampling-50 years after Shannon,” Proc. IEEE, vol.88, no.4, pp.569–587, April 2000.