社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

言語的情報と非言語的情報の音響的分離と音声コミュニケーション

峯松　信明†

† 東京大学大学院情報理工学系研究科，〒 113-0033 東京都文京区本郷 7-3-1
E-mail: mine@gavo.t.u-tokyo.ac.jp

あらまし　音声に含まれる音響的特徴の分離モデルとして，音声の生成過程に着眼し，音源と声道の特性に分離するソース・フィルターモデルが従来より広く使われている。しかし，声道形状を音響的に表象するスペクトル包絡特性は，本来独立な情報と考えられる音声の言語的情報と非言語的情報の両方に関与する。そのため，例えば，音声認識における不特定話者音響モデルは，同一の言語的情報を担う音声を大多数の話者から集め，統計的に音響モデリングを行うことが一般的である。本稿の前半では，幼児の言語獲得における音声模倣行為，音声コミュニケーションの獲得に困難を示す自閉症者の音声模倣行為，更には動物における音声模倣行為などを概観し，音声の中に同居する言語的情報を担う特徴と非言語的情報を担う特徴とを音響的に分離するモデルの必要性を主張する。また，情報分離が不完全な音響モデリングは，音声コミュニケーション能力の計算機上での実現を目的とした音響モデリングではなく，声帯模写能力の計算機上での実現を目的とした音響モデリングとして解釈すべきであること，そして，不完全な情報分離のみでは，本来，人間であっても音声コミュニケーションは困難となることについて言及する。本稿の後半では，二種類の情報の分離を目的として筆者が近年提案している音声の構造的表象について触れ，幾つかの実験結果を紹介する。
キーワード　ソース・フィルターモデル，言語的・非言語的情報，スペクトル包絡特性，音声コミュニケーション，音声模倣と声帯模写，自閉症，音声の構造的表象，f-divergence に基づく完全変換不変性

Acoustic separation between linguistic and extra-linguistic information in speech and its significant importance to enable speech communication

Nobuaki MINEMATSU†

† Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
E-mail: mine@gavo.t.u-tokyo.ac.jp

Abstract　The source-filter model, which was derived from observations of speech production, has been widely used to separate speech features into two parts: vocal source characteristics and vocal tract characteristics. However, the latter characteristics, often called as spectrum envelopes, transmit both linguistic information and extra-linguistic information, which are intrinsically independent of each other. This is why a speaker-independent acoustic model of a linguistic content for ASR is often built statistically by collecting utterances of that linguistic content from a large number of speakers. In the beginning part of this paper, after reviewing infants' vocal imitation for language acquisition, the vocal imitation observed in severely impaired autistics who have difficulty in speech communication, and the vocal imitation of animals, we claim the importance to derive the acoustic modeling which can separate acoustic features for linguistic information and those for extra-linguistic information. We also insist that the acoustic modeling with incomplete separation should be suited not for realizing speech communication ability on machines but only for realizing impersonation ability on machines. Further, we describe that, only with incomplete separation, speech communication has to be difficult even for humans. In the ending part of this paper, we introduce the structural representation of speech, which we proposed to realize the information separation for creating human-like machines, and show some experimental results obtained by using the proposed representation.
Key words　source-filter model, linguistic and extra-linguistic information, spectral envelope, speech communication, vocal imitation and impersonation, autism, speech structure, transform-invariance based on f-divergence

1. はじめに

音声は，時間も振幅も連続的な値を有する一次元信号（波形）として観測される。それを標本化・量子化して整数値列とし，計算機上で各種の処理が行われる。計算機にとって音声とは単なる数値列でしかないが，この数値列の中に様々な情報が埋め

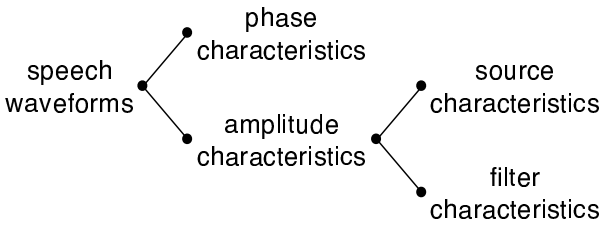

図 1　二段階の分離に基づく特徴量抽出

込まれて（符号化されて）いる。そして人間はその情報をいとも簡単に解読して（復号化して）しまう。多様な情報を適切に反映しつつ数値列を導出するのが音声合成であり，その数値列から多様な情報を的確に抽出するのが音声認識・理解である。

これらの技術を構築する場合，数値列のある側面を切り落とし，処理の効率化を図っている。例えば人間の聴覚は音声信号の位相成分には鈍感であるとの知見から，パワースペクトルのみを特徴量として使用する場合が多い。更に，音声生成を「声帯による音源生成」と「声道による共鳴」との二段階に分け（ソース・フィルタモデル），両者による音響特性を関数の積で表現することで，後者による音響効果のみに着眼することも多い。現在の音声認識技術が好例であり，音源の音響特性を切り落としたスペクトル包絡特性を基本的な音響特徴量として用いている（**図 1** 参照）。二段階の分離を通して得られる包絡特性であるが，尚，様々な情報源がこの音響量を変形させる。

音声に含まれる情報は大きく言語情報，非言語情報に分類される（言語情報は文字面情報だけに限定した言語情報と，文字面では表現困難なパラ言語情報に細分化される）。スペクトル包絡（共鳴特性）は声道形状を直接反映した音響量であるが，言語，パラ言語，非言語情報の何れによっても容易に変形を被る。

不特定話者単語音声認識，テキスト非依存話者認識を例として考える。包絡特性 o は，単語 w（言語情報），話者 s（非言語情報），何れにも依存する。統計的音響モデルの構築を考えた場合，単語認識の場合は $P(o|w)$ を，話者認識の場合は $P(o|s)$ を推定することになる。ここで，認識対象とは独立な要因を期待値（周辺化）操作で消失させることが広く行われている。

$$P(o|w) = \sum_s P(o|w,s)P(s|w) \approx \sum_s P(o|w,s)P(s)$$

$$P(o|s) = \sum_w P(o|w,s)P(w|s) \approx \sum_w P(o|w,s)P(w)$$

しかし言語情報（単語）と非言語情報（話者）は，そもそも独立した情報である。にも拘わらず，それらを運ぶ音響的対象物（特徴量）として，各々に対応した特徴量（o_w や o_s）を求めずに，共通項 $P(o|s,w)$ に対する期待値操作で各々の音響モデルを導出する。音声工学では常套手段となっているが，機械学習を専門とし，音声・話者認識は一応用として捉えている研究者には，これを不思議な方法論と考える研究者もいる [1]。

本稿では，人間のように汎化能力の高い音声情報処理の計算機実装を目的とした場合，欠けている基礎技術として「音声に含まれる言語的情報を，非言語的情報から音響的に分離して抽出する技術」を主張する。本稿の前半では，この主張を裏付ける発達心理学，発達障害学，更には進化心理学における各種知見について触れる。後半では，この主張に対する技術的回答として筆者が近年提唱している音声の構造的表象に触れ，幾つかの実験結果を紹介し，最後に本稿をまとめる。

2. 発達的・進化的側面から考察する音響モデリングの技術的欠損

2.1 発達的側面から考察する技術的欠損

幼児の言語獲得は「音声模倣・学習」を基本とする [2]。他個体の発声を積極的に模倣する行為である。ここで注意すべきは，彼らの模倣の音響的対象物である。幼児の音声模倣は，音響的模倣（声帯模写）では無い。音響的には何を真似ているのか？

「親の声をシンボル（音韻，平仮名）列に変換し，個々のシンボルを自らの口で生成する」という説明は不適切である。彼らは音韻意識が未熟であり,「しり取り」も困難な状況にある [3]。発達心理学の文献を調査すると，各種用語でこの模倣対象を説明している。[4] では「幼児は単語全体の語形・音形を獲得し，その後，個々の分節音を獲得する」と説明し, [5] では「語形の全体ゲシュタルトを認知する」と述べ, [6] では「related spectral pattern」と呼んでいる。これらは同一対象と解釈できるため，本稿では「語ゲシュタルト」と呼ぶこととする。

語ゲシュタルトに話者情報（非言語的特徴）が含まれていれば，幼児は音響的模倣を試みることになる。彼らは親の声と自らの声の音響的ずれに無頓着な模倣を行っており，語ゲシュタルトは音声から話者情報が切り離された音響パターンと言える。著者は国内外の発達心理学研究者に，語ゲシュタルトの物理的定義の提示を促したが，明確な回答はなかった。

さて「幼児の聞く声の大半は両親の声であり，また，自らが話せるようになると，その子の聞く声の約半分は自らの声である」という記述を否定することは困難である。即ち，人が聴取する音声の話者性は極めて偏りが大きい。そして，この話者的に偏った音声の聴取を通して，頑健な情報処理を獲得する。話者情報を切り落として言語的情報のみを音響パターンとして抽出する能力があれば，当然の帰結である。その一方，話者情報の分離技術が確立せず，集めることで $P(o|w)$ を推定する枠組みを採択すれば，話者バランスがとれた音声サンプルが必要になる。かつて IBM 社が自社製の音声認識エンジンの宣伝に用いた「集めた話者数」は，35 万人であった。

音声模倣が音響的模倣になる場合があるのだろうか？そのような事例は（重度）自閉症者に見られる。七色の声を持つと呼ばれる声優の中村メイ子の声をそっくり真似る例 [7]，外国語発音練習やカラオケにおいて，音響的模倣以外の真似方が難しい例 [8]，相手そっくりの声を模倣する例 [9]〜[11]，音声に限らず車や列車の音など，様々な音響音を模倣する例は，自閉症関連図書において頻出する。刺激音をそのまま記憶し（そのため音響的汎化能力も低下すると考えられる），再生しようとする情報処理が主体となっている訳だが，重度自閉症者の場合「音声コミュニケーションが困難となる場合が多い」という事実は注目に値する。中には，母親の音声は正しく認識・理解できるが，母親以外の音声への対応が難しい例もある [12]。

ある話者の音声を学習データとして音声合成システムを構築すれば，任意のテキストをその話者の声で読み上げる（所謂，声帯模写）。成人話者を多数集めて構築した音響モデルで子供の音声を認識すれば，認識率は下落する。音声合成，認識ともに，言語情報と非言語情報が同居したままスペクトル包絡特性をモデル化する点では同じである。つまり，現在の音響モデリング技術は音そのものを記憶・モデル化対象としており，これは，(重度）自閉症者の情報処理と類似していると言える。環境によって多様に変形する音声を媒体としたコミュニケーションに困難を抱える事実を考えると，(典型的発達を遂げた）人間が有する高度な音響的汎可能力の実現に必要な基礎技術は，言語的情報と非言語的情報の音響的な分離であると考える。

2.2 進化的側面から考察する技術的欠損

動物を対象とした場合，音声模倣は稀な行為と位置づけられている。例えば霊長類では，人のみが行う行為であると考えられている [13]。霊長類以外で音声模倣を行う種は鳥，クジラ，イルカなどが確認されているが [14]，動物の音声模倣は音響的模倣が基本となっている [14]。また，進化人類学の実験によれば，人以外の霊長類は相対音感が乏しく，移調前後のメロディーの同一性判定が困難であることが示されている [15], [16]（注1）。即ち，人以外の霊長類は極端な絶対音感を有している（注2）。

アスペルガー症候群（自閉症の一種）を患う者として世界で初めて書籍を出版した [9] グランディンは動物学の教授であるが，彼女は，自閉症者と動物の情報処理における類似性を指摘している [17]。いずれも，入力刺激の詳細な様子をそのまま記憶・保持する傾向が強い。入力された情報を無意識的に取捨選択できず，汎化能力に乏しく，情報過多の渦に巻き込まれる様子は多くの自閉症関連図書に散見される [8], [11], [18], [19]（注3）。自閉症者の多くは絶対音感保有者である [20]。

以上，人間と動物の音情報処理の差異に関して，筆者の文献調査の結果を述べた。音を用いた情報伝達を行う場合，情報の同一性を保証するために，音響的同一性が必要とされるのか否か，が問うべき焦点であると考える。必要であれば，音響的に同一の音を自らが生成したり，他者に要求することになる。重度自閉症者や動物の音声模倣，更には，動物における移調前後のメロディー同一性認知の欠損などはその良い例である。

音声認識における音響的照合とは，ある発声と別の発声（あるいは音響モデル）の言語的な同一性検証を，音響的な同一性検証を通して行う技術である。言い換えれば，二種類の同一性を置換可能と仮定して，初めて成立する技術である。この仮定は正しいのだろうか？身長が 2.5m 近い世界一の巨人と 1.0m に満たない世界一の小人が難なく会話する様子が TV で報道されることがある。世界一の音色・声色の音響的ずれを持つ両者は，それを全く気にせず会話を楽しむのである。

筆者はこの二種類の同一性は置換可能なものではないと考える。「(言語的）情報の同一性を保証する場合に，音響的同一性を必要としなくなった」種が人間である，と考えている。換言すれば「音のある側面が同一であれば，(その側面に関係する）情報の同一性を認知できるようになった」種が人間である，と考えている。前節同様，本節においても，非言語的情報に非依存な音響パターンを通して言語情報の同一性を検証する技術の構築が，高い音響的汎化能力の実現には必要であると考える。

(注1)：但し，1 オクターブずらすと同一性が分かるとのことである。
(注2)：彼らがメロディーを音名で記述できたり，採譜できる訳では無い。違う音は違う音，と認識しているだけである。
(注3)：ある当事者は，自閉症とは「情報の便秘」である，と述べている [11]。

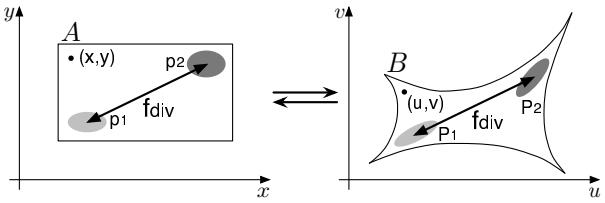

図 2　連続かつ可逆な変形に対して不変な f-divergence

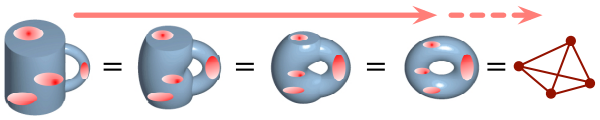
図 3　f-divergence に基づく形態的不変性

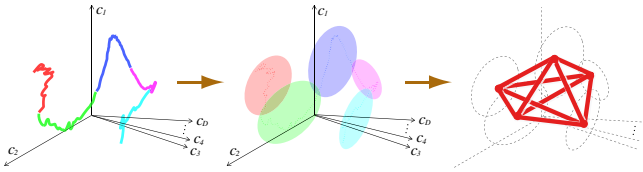
図 4　f-divergence を用いた一発声の構造化

3. 写像不変量に基づく構造的表象とその応用

3.1 完全なる写像不変量とそれに基づく音声表象

声質変換・話者変換に代表されるメディアモーフィングは，特徴量空間の写像として一般化できる。非言語的情報の違い（例えば話者の違い）による特徴量変形は，ある写像となる。よって，非言語的情報の違いに非依存・独立な音響パターンとは，写像不変量のみで構成されたパターンとして導出できる。

事象が分布として記述される特徴量空間を考える。t>0 において凸な関数 $g(t)$ を用いて次式で定義される f-divergence は，連続・可逆な全写像に対して不変となる（十分性，**図 2**）。更に 2 事象に関する量を $\oint M(p_1(\boldsymbol{x}), p_2(\boldsymbol{x}))d\boldsymbol{x}$ で定義すると，これを写像不変とするのは f-div. のみである（必要性）[21]。

$$f_{div}(p_1, p_2) = \oint p_2(\boldsymbol{x}) g\left(\frac{p_1(\boldsymbol{x})}{p_2(\boldsymbol{x})}\right) d\boldsymbol{x}$$

空間内に事象が N(>2) 個存在すれば，${}_NC_2$ 個計算される事象間の f-div. は不変であり，距離行列は不変となる。一般に距離行列は一つの幾何学的形態を規定するため，この距離行列を，事象群に対する（写像不変な）構造的表象と呼んでいる [22]。例えば**図 3** に示す連続・可逆な空間写像による形状の変形を考える。変形された形状の表面及び内部に分布としての事象群が存在している場合，f-div. は如何なる写像によっても変化しないため，f-div. を用いて計算される距離行列は不変となる [23]。

この f-div. を用いて一発声を構造化する様子を**図 4** に示す。特徴空間における発声軌跡を有限個の分布列に変換し，全ての分布間距離を f-div. で計測し，構造的表象を導出している。

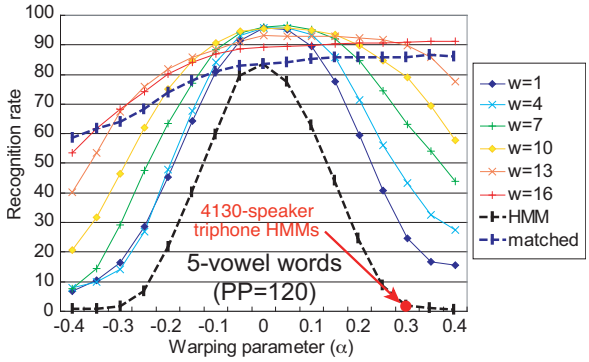

図 5　五母音単語セットに対する認識率

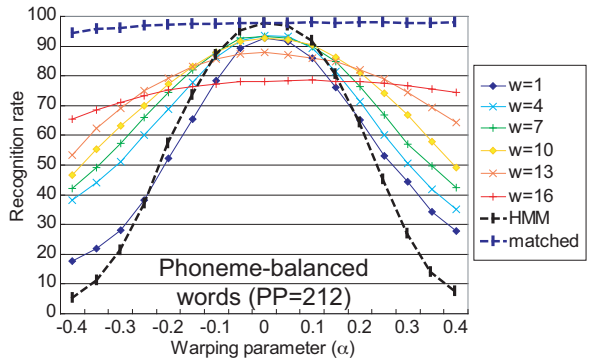

図 6　バランス単語セットに対する認識率

3.2　音声の構造的表象の応用例

構造的表象に基づく音響的照合は，話者適応を施した後の音響スコアを，話者適応を明示的に施すことなく算出可能である [23]。そのため，例えば音声認識において，話者性に起因するミスマッチに対する極めて高い頑健性が示されている [23]。日本語五母音を並べ替えて構成できる 120 単語，及び音韻バランスのとれた 212 単語を対象にした孤立単語音声認識結果を**図 5**，**図 6** に示す。なお，評価用音声に対して周波数ウォーピングを施すことで，巨人や小人に相当する音声を人工的に作成し，使用している。横軸（α）の負（正）は，声道長の長（短）に対応し，$|\alpha|$=0.4 で約倍，半分になる。図中 HMM とは，比較実験として行った単語 HMM の性能であり，matched とは，α の各値に対応した学習データを用いた単語 HMM の性能である。w は部分空間化[注4]における次元幅である。

母音単語の場合，適切な w を設定することで，構造的表象は matched HMM と同等の性能を示している。これは，話者性の違いが完全にキャンセルされていることを示している。バランス単語においても高い頑健性は示されているが（w=10, 13 など），matched には及んでいない。一部の子音（無声破裂音，摩擦音など）は話者差による変形が母音に比べて小さいため，音と音の相対的な関係だけでは十分に対応できていないと解釈される。音と音の関係性に基づく情報処理と，音の絶対的な特性に基づく情報処理の融合については [24] を参照して戴きたい。

4.　音高の相対音感の一般化としての構造的表象

4.1　音高の相対音感の一般化

本節では，構造的表象とメロディーの相対音感との類似性について記述する。相対音感に基づいてメロディーを階名で書き起こす場合，調を上下させても（移調），書き起こされる階名列（ドレミ列）は変わらない。調に不変な音同定が行われる。構造的表象は空間内の事象間距離を不変に計測する手段を与えるが，階名書き起こしも**図 7** に示す，音高差（音程）が，移調前後で不変であることが必要十分条件である [25], [26]。移調しても，適切に黒鍵を使うことで，音高差の不変性は保たれる。

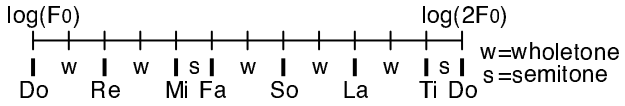

図 7　長調におけるオクターブ内の音配置

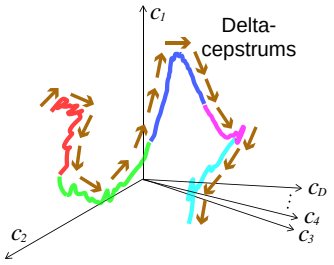

図 8　Δ ケプストラム

楽器音は，基本周波数軸に対して離散的に存在するが，音声のイントネーションのように連続的に変化する音高変化パターンに対して音高差の不変性を考えれば，それは ΔF_0を用いた F_0 パターン表象に相当する。声帯振動の個人差に起因する静的バイアスが除去されるため，不変的な F_0パターン表象となる。

同様に Δ ケプストラム（以降 $\Delta \boldsymbol{c}$）による発声表象を考えた場合（**図 8** 参照），音響機器特性の違いは $\boldsymbol{c}'=\boldsymbol{c}+\boldsymbol{b}$ と定ベクトル $\boldsymbol{b}$ を足す演算であるため不変的となるが，声道長の違いによる影響を強く受ける。声道長の差異による音響変化は $\boldsymbol{c}'=A\boldsymbol{c}$ で近似できるが [27]，この時行列 A は，回転性の非常に高い行列となる [28]。即ち，特徴空間において軌跡が回転してしまう（軌跡の方向がバイアスとしての物理的意味を持つようになる）。F_0 は一次元の物理量であるため，特徴量空間では正・負の方向性を持つだけで，連続的な回転操作は不可能である。しかし，特徴量次元が増えると「軌跡の方向」がバイアスとして物理的意味を持つようになり，そのため，構造的表象では距離だけを用いた表象となっている。写像による変形がシフト，回転だけであれば事象を点で表象し，ユークリッド距離を用いた距離行列が不変表象となるが，シフト，回転以外の変形操作（**図 2** 参照）に対しても不変な表象とするためには事象を分布化し，f-div. に基づく距離行列を計算することが必要になる。

なお筆者は，音声の構造的表象を話者（非言語的情報）正規化手法とは考えていない。本来正規化とはある変数（例えば声道長）が基準値（例えば 17cm）をとるように観測量を変形する操作を言う。構造的表象は上記したように，静的バイアスによって変形する特徴量を分離した後に残る特徴量であり，非言語的な情報を分離した後に残る特徴量であると考えている[注5]。

(注4)：部分空間化については [23] を参照して戴きたい。

(注5)：スペクトルから調波構造を取り除いて包絡化する操作を，スペクトルのピッチ正規化とは言わないことと同じである。ピッチの情報を取り除いている。

4.2 韻律的特徴としての音声の構造的表象

以上，構造的表象が一次元物理量の F_0 を対象とした相対音感の自然な多次元化・一般化であることを示した。F_0 パターンは F_0（一次元）の動的変化パターンであり，相対音感はそれを不変的に認知するための情報処理である。**図 8** に示した軌跡は，音色（スペクトル包絡，多次元）の動的変化パターンであり，構造的表象はそれを（相対音感的に）不変的に扱うための枠組みである。前者は F_0 の，後者は音色の動的変化パターンでしかない。前者は韻律的特徴と呼ばれるが，筆者は後者も韻律的特徴(注6)として考えている。音色を分節的特徴と呼ぶのは，音声を音韻（セグメント）列として表象することが前提である。**第 2.1 節**に示したように，音韻意識が未発達の幼児が親の歌や発声を模倣する際に，F_0 パターンをドレミ列（シンボル列）として解釈した後に模倣したり，音色パターンを音韻列（シンボル列）として解釈した後に模倣しているとは筆者は考えていない。F_0 や音色の動的変化パターンから，ある側面だけを自らの発声の中で再度再生しているだけであると考えている。

絶対音感を持つと，相手のハミングを復唱する際に，一旦ドレミ列（但し，音名列）として解釈した後に復唱する以外の復唱の仕方が想像できない者もいる。音名としての同定が不可避的に行われるからである。通常言語を獲得すると，提示音声に対して音韻列を知覚できるようになるが，成人になってもこれに困難を示す場合がある（音韻性 Dyslexia [30]）。音韻意識が未熟な幼児（や音韻性 Dyslexia の方々）による相手の発声の模倣行為を考えた場合，模倣対象となっている「声のある側面」に対して，分節的特徴という言葉を使うのは不適切であろう。

5. 非可観測な対象のモデル化に対する考察

5.1 不可避的な音響的バイアスと非可観測性

音声のスペクトル包絡は体格，年齢，性別などに起因する音響的バイアスによって不可避的に影響を被るが，この影響を取り除いた後に残るものとして音声の構造的表象を考えている。言い換えれば，言語的（パラ言語的）情報が同一であれば，誰の声にでも観測される音声の共通項をあぶり出すことに相当し，必然的に，抽象的な音響的表象となる。そして，この共通項は（常に音響的バイアスを被るという意味で）直接的には観測することは，論理的に不可能である。音声がある個人の調音器官を通して生成される以上，この音響的バイアスは不可避である。本節では，この不可観測な対象をモデル化することの意義について考察する(注7)。直接的に観測できない以上，モデルパラメータの精度を，誤差の大小などの基準で評価することは出来ない（例えば構造的表象で言えば，推定されたエッジ長の誤差を直接的に評価することは出来ない）。可能なのは，その抽象的なモデルから間接的に生成される信号（例えば構造からの音声合成 [31] によって生成される音声信号）での誤差に着眼したり，或は，ASR や CALL 応用のように，そのモデルを応用することの効果でもって評価する [23], [32] こととなる。このように，直接的に観測すること・評価することが原理的・論理的に不可能な対象を積極的にモデル化することの意義は何であろうか？例えばソース・フィルタモデルによって得られるスペクトル包絡(注8)に対するモデルでは満足せずに，その裏側にあるかのように思われる対象物を，積極的に検討対象とする意義は何であろうか？(注9) ここでは**第 2. 節**の文脈に沿って再考する。

5.2 観測量の直接的・絶対的記憶と情報分離

第 2. 節にて自閉症者の情報処理について幾つか記述した。筆者は自閉症研究を専門とする立場には無いため，彼らの情報処理についてその詳細を述べることは出来ない。しかし，彼ら自らが執筆する書籍や論文，更には自閉症を研究対象とする研究者による書籍や論文を調査する限りにおいて，刺激の物理的側面（感覚センサーの出力）をそのまま記憶する傾向が非常に強いことは否定することが出来ない事実だと考えている。

例えばグランディンは「自閉症者は感覚野メインで生きている」と述べている。つまり，その後の連合野での処理が典型的発達を遂げた一般人と比較すると弱い，ということである。**第 2. 節**では自閉症者や動物の情報処理についての各種文献調査の結果を述べたが，刺激の物理的側面を直接記憶する（絶対的記憶）と，多様に変化する環境に対する頑健性が乏しくなるのは論理的必然である。**第 2. 節**では音声模倣という現象に焦点を当てて記述したが，同様の議論は他の現象でも可能である。興味のある読者は各種文献を参照して戴きたい。

このように，物理的に直接観測ができるものを対象としたモデル（ある音韻に対応するケプストラムベクトルを複数集め，それをガウス分布でモデリングするなど）は，当然，環境によって異なる音響的バイアスの影響を直接的に受ける。技術的には，音響的ミスマッチを避けるために事前にモデルを修正することで対処できる。しかし，健常者の音声模倣を考えると，彼らは音響的ミスマッチに対して極めて鈍感である。

第 2. 節で述べた様に，音声の技術構築においては，種々の知見に基づき，情報を適宜捨象する形で特徴抽出，音響モデリングを試みてきた。位相スペクトルの切り落とし，調波構造の切り落としがそれに相当する。しかし，人の聴覚が位相スペクトルに鈍感なように，人は言語を獲得するときに，親の音声の非言語的情報には鈍感である。確かに，人は音声の非言語的情報を認知し，それに基づいた行動をとることができる。その意味では敏感である。しかし，言語獲得（音声模倣）においては，

(注6)：藤崎は prosody を次のように定義している [29]。Prosody is defined as the systematic organization of individual linguistic units into an utterance, or a coherent group of utterances, in the process of speech production. Its realization involves both segmental and supra segmental features of speech, and is influenced, not only by linguistic information, but also by para-linguistic and non-linguistic information. **図 4** に示した systematic organization は prosody そのものである。

(注7)：ここで言う不可観測性は，例えば隠れマルコフモデルにおける隠れ状態とは意味が異なる。隠れ状態や隠れ変数における非可観測性とは，決定論的に扱うことができない性質（確率論的にのみ扱うことができる性質）を指す。

(注8)：厳密には有声音のスペクトルは常に調波構造を伴うので，包絡特性は直接観測できる対象ではない。しかしここでは，包絡特性を可観測な対象として記述している。音声認識の世界で包絡特性を observation と呼ぶという意味において，可観測量として記述している。

(注9)：一般論としては，モデルパラメータが非可観測であることは特殊なことでは無い。

自らの発声にその情報を反映しようとしない。その意味で鈍感である。筆者はこの鈍感さ（無視できる能力）こそが，頑健な情報処理を可能にしていると考えている。そしてその能力の技術的実装が，言語的情報と非言語的情報の音響的分離となる。

脳科学では，感覚器から入力された情報が一旦分離される情報処理モデルが広く受け入れられている。視覚情報の場合は，第一次視覚野からの情報が腹側経路と背側経路とに分かれ，各々が，what の情報，where（あるいは how）の情報を表象する [33]。聴覚情報の場合でもこれに倣い，言語情報，非言語情報の分離モデルを検討している研究もある [34], [35]。これらの研究動向から見ても（実装方式の是非は問えないが），情報を分離する技術を構築することは，より人間らしい情報処理の実装に近づいていると筆者は考えている [36]。

6. ま と め

本稿では，音声システムの頑健性の向上に主眼を置き，現在の音響モデリング技術に欠損している基礎技術として「言語的情報と非言語的情報の音響的分離」について，各種の情報源を参照する形で考察した。期待値操作（周辺化操作）による統計モデリングや，適応処理によるモデル修正は技術的には非常に有効なツールであるが，筆者は，これだけでは人間の高い汎化能力の実現は不十分であると考えている。本稿では，自閉症や動物の（特に音に対する）情報処理について各種分野で得られた知見を網羅することで,「言語的情報と非言語的情報の音響的分離」を主張し，筆者らが近年検討している音声の構造的表象についても紹介した。見えるもの，聞こえるものを直接対象とするのではなく，その裏側に存在すると思われるものをモデル化する場合，当然，理論的考察が必要となる。人間の高い汎化能力の実現には，今後，様々な理論の登場が必要となるのかもしれない。なお，本稿で示した各種考察をより詳細に記述した論文が [36] である。こちらも是非参照して戴きたい。

文　献

[1] N.D. Lawrence *et al.*, “Dealing with high dimensional data with dimensionality reduction,” Tutorial of INTERSPEECH, 2009
[2] P. K. Kuhl, *Nature Reviews Neuroscience*, 5, 831–843, 2004
[3] 原，コミュニケーション障害学，20，2，98–102，2003
[4] 加藤，コミュニケーション障害学，20，2，84–85，2003
[5] 早川，月刊言語，35，9，62–67, 2006
[6] P. Lieberman, *Child Phonology vol.1*, Academic Press, 1980
[7] 深見，ひろしくんの本 (V)，中川書店，2006
[8] 綾屋他，発達障害当事者研究，医学書院，2008（但し，著者らの対談を含む）
[9] T. Grandin，我, 自閉症に生まれて，学研，1994
[10] L.H. Willey，アスペルガー的人生，東京書籍，2002
[11] ニキリンコ，スルーできない脳〜自閉は情報の便秘です〜，生活書院，2008
[12] 東田他，この地球にすんでいる僕の仲間たちへ，エスコアール，2005
[13] W. Gruhn, *In: Proc. Int. Conf. on language and music as cognitive systems*, 2006
[14] 岡ノ谷，春季音講論，1-7-15，1555–1556，2008（質疑応答含む）
[15] A.A. Write *et al.*, *J. Exp. Psychol. Gen.* 129, 291–307, 2000
[16] M.D. Hauser *et al.*, *Nature neurosciences*, 6, 663–668, 2003
[17] T. Grandin，動物感覚〜アニマル・マインドを読み解く，日本放送出版協会，2006
[18] 泉，僕の妻はエイリアン，新潮社，2005
[19] 藤井他，自閉症，新曜社，2007
[20] U. Frith，自閉症の謎を解き明かす，東京書籍，1991
[21] Y. Qiao *et al.*, IEEE Trans. on Signal Processing, 58, 7, 3884–3890, 2010
[22] 峯松他，信学技報，SP2005-12，1–8，2005
[23] N. Minematsu *et al.*, Journal of New Generation Computing, 28, 3, 299–319, 2010
[24] Y. Qiao *et al.*, *Proc. IEEE Workshop on Automatic Speech Recognition & Understanding*, 118–123, 2009
[25] 谷口，音は心の中で音楽になる，北大路書房，2003
[26] 東川，読譜力－「移動ド」教育システムに学ぶ，春秋社，2005
[27] M. Pitz *et al.*, *IEEE Trans. SAP*, 13, 5, 930–944, 2005
[28] D. Saito *et al. Proc. ICASSP*, 4485–4488, 2008
[29] H. Fujisaki, “The command-response model for F_0 contour generation and its implications in phonetics and phonology,” Keynote lecture in The 8-th Phonetics Conference of China, 2008
[30] S. Shaywitz, 読み書き障害（ディスレクシア）のすべて〜頭はいいのに本が読めない〜, PHP 研究所, 2006
[31] D. Saito *et al.*, *Proc. INTERSPEECH*, 2047–2050, 2009
[32] M. Suzuki *et al.*, *Proc. Int. Workshop on Automatic Speech Recognition and Understanding*, 574–579, 2009
[33] L. G. Ungerleider, “Two cortical visual systems,” in *Analysis of Visual Behavior* (edited by David J. Ingle), 549–586, MIT Press, 1982
[34] S.K. Scott *et al.*, *Trends in Neurosciences*, 26, 2, 100–107, 2003
[35] P. Belin *et al.*, *Nature neuroscience*, 3, 10, 965–966, 2000
[36] 峯松他，“音声に含まれる言語的情報を非言語的情報から音響的に分離して抽出する手法の提案　〜人間らしい音声情報処理の実現に向けた一検討〜”，電子情報通信学会論文誌，J94-D，1，2011