音声の構造的表象を通して考察する幼児の音声模倣と言語獲得

Consideration on infants' speech mimicking and their language acquisition based on the structural representation of speech

峯松信明 †，西村多寿子 ‡，櫻庭京子 ∗

Nobuaki Minematsu†, Tazuko Nishimura‡, Kyoko Sakuraba∗

† 東京大学大学院新領域創成科学研究科 / Graduate School of Frontier Sciences, The University of Tokyo

‡ 東京大学大学院医学系研究科 / Graduate School of Medicine, The University of Tokyo

∗ 清瀬市障害者福祉センター / Kiyose-shi Welfare Center for the Handicapped

mine@gavo.t.u-tokyo.ac.jp, nt-tazuko@ams.odn.ne.jp, sakuraba@mtd.biglobe.ne.jp

Abstract

In speech communication, acoustic distortions are inevitably involved by speakers, channels, and hearers. However, infants acquire a spoken language mainly with speech samples of their mothers and fathers. They can solve the variability problem only with a remarkably biased speech corpus. Why and how is it possible? To answer this hard question, we already proposed a speaker-invariant structural representation of speech. In this report, the proposed representation is mathematically shown to be invariant also with non-linear transformations. Based on this representation, the speech recognition processes of dyslexics and autistics, often viewed as paradox, could be taken for granted. Finally, we discuss that speech communication should be based on relative sense of sounds.

1 はじめに

音声コミュニケーションには，話者・環境・聴取者に起因する，多様な音響歪みが不可避的に混入する。その一方で幼児は，大部分が「母親と父親の音声」という音声資料の提示を通して音声言語を獲得する。これは，音響的に非常に偏った話者性の音声資料の提示を通して，多様な音響歪みに関する対処法を獲得することを意味する。偏った音声資料の提示は，その後一生続く。何故ならば，人の聞く声の半分は自分の声だからである。人は偏った音声提示しか受けられないのである。何故，このように音響的に偏った音声提示環境の下で，人（幼児を含む）は多様な音響歪みに対処できるのだろうか？音響音声学／音声工学では，この多様性問題を直接的に解くことはせず，個々の音素の音響モデルを，数千・数万という話者の音声を集め，分布としてモデル化することでその解決を図ってきた。それでも多様性問題は解けず，音響モデル適応／特徴量正規化などの技術を編み出して来た。全く異なる戦略を示す両者の，本質的差異はどこにあるのだろうか？

話者 A の音声を書き起こす。話者 B の音声を書き起こす。この時，話者 A によって発せられたある音響事象を「あ」という記号で表記し，話者 B によって発せられたある音響事象も同様に「あ」という記号で表記する。当然，両音響事象間に物理的等価性は保証されない。物理的に異なる音響事象群を，同一の表記を用いて書き起こす訳である。なぜ話者毎に「あ」という記号の変種を用意することなく,「あ」と表記できるのだろうか？

提示された曲を，階名を用いて「ドレミ」として書き起こす。曲が階名として聞こえてくる聴取者は，その曲を移調しても，書き起こされる「ドレミ」列は変わらない。相対音感者である[1]。移調によっていくら「ド」の音高が変わろうとも，彼らは「ド」として表記する。第一著者は絶対音感を持っており，この階名での書き起こしが全くもって理解できない一人である。異なる音高に同一の音ラベルを振ることなど，全く理解不能である。

異なる話者間で「あ」の同一性が感覚できない人がいるのだろうか？音の絶対特性に執着し，両者の同一性が感覚できない人がいるのだろうか？感覚できない「機械」が，限られた話者の音声から構築された（特定話者）音声認識器である。そして，感覚できない「人」として，一部の自閉症者がいる[1]。優れた絶対音感を持つ率が，健常者と比較して遥かに高い自閉症者の中には,「ハ」の音（固定ドとしてのドレミの場合は「ド」の音）で始まらない「カエルの歌」を，それと認めない者もいる[2]。

相対音感者による階名による書き起こしは，音階の構造（全全半全全全半という音高遷移の枠組み）をメロディーの中に感覚し，例えば長調の場合，主音をドとして認識し，同様にして，上主音，中音，下属音を，レミファ，として認識する。即ち，音列の流れを通して，全体的なメロディー構造（音楽学では，これを「横の構造」と言う）

[1] なお，階名の書き起こしが出来ない（ハミングしかできない）相対音感者もいる。この場合，彼らは「言語化が困難な」相対音感者である。

の認知が先に起こり，それに基づいて個々の要素音の（他音群との関係によって定まる）機能的・相対的価値を認識する訳である。その結果，要素音の絶対的物理特性とは全く独立に，個々の要素音が同定されることになる。物理的には全く異なる二音が同一の機能的価値を有した時，両者が同一音として「聞こえる」ことになる[3, 4]。

幼児は，極端に限られた音声の多様性に接することで，広範囲に渡る音声の多様性に対処できるようになる。発達心理学によれば，「幼児の音声言語獲得は，分節音の獲得の前に語全体の音形・語ゲシュタルトの獲得から始まる」とされている[5, 6]。個々の音韻意識が定着するのは小学校入学以降であり，それまでは「しりとり」に難を示す児童もいる[7]。即ち幼児の音声コミュニケーション（例えば音声生成）は，個々の音韻（モーラ）を一つ一つ音に変換する形では（少なくとも意識の上では）行なうことは困難である。日本語には母音が5つあり，それが/あ/い/う/え/お/であることを知る以前に，幼児は両親と音声コミュニケーションを行ない，自己主張までする。

全体的なメロディー構造を通して，移調された曲同士の同一性を感覚し，更には，個々の要素音の（階名としての）同一性を感覚する。その結果，物理的に全く異なる二音を同一であると感覚する。幼児の言語獲得も同様に，語全体の音形の獲得から開始される，とした場合に，この語全体の音響表象が，音楽同様，音声の移調（非言語要因による不可避的な音響変動）による多様性問題を解く鍵になるのだろうか?従来より筆者らは，非言語的要因による音響変動に対して簡素な数学モデルを考え，この問題を解決してきた[8, 9, 10, 11, 12, 13]。本稿ではこれを一般化し，非常に広範囲な変換（非線形変換を含む）においても，移調不変な構造的表象が普遍的に存在することを数学的に示す。そして幼児の音声模倣，更には，自閉症・失読症者の音声認知を，この構造的表象を通して考察する。最後に，自閉症者のビヘービアと音声認識システムのそれとが非常に類似していることを指摘すると共に，人間の音声情報処理と機械の上に実装された情報処理との本質的・根源的な差異について，音情報処理の生物進化に伴う変遷を踏まえ，筆者らの意見を述べる。

2　非言語的音響変動不変の音声の構造的表象

2.1　2つの空間における頑健な不変量

図 1 に示す様な，二つの空間 A と B を考える。両者には一対一の対応関係があり，空間 A のある点は空間 B の対応点へ写像され，逆もまた成立する。但し，その写像関数は明示的には与えられていないとする。以下，一般性を失わない範囲で2次元空間を用いて説明する。空間 A，B の対応する二点を (x, y)，(u, v) とし，両空間の対応付け（変数変換）を一般的に下記の様に示す。

$$x = x(u, v), \quad y = y(u, v) \tag{1}$$

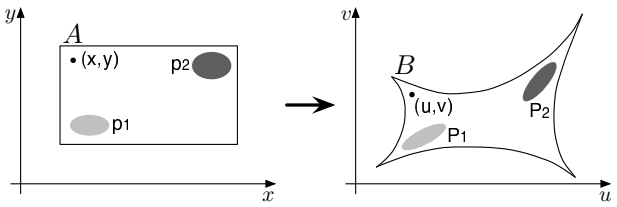

Figure 1: 一対一対応関係を有する二つの空間 A と B

空間 A における事象を考える。但し，全ての事象は空間内の点ではなく，確率密度分布関数として存在するものとする。即ち事象 p は次式を満たす。

$$1.0 = \iint p(x, y)dxdy \tag{2}$$

空間 A における積分演算は，変数変換によって空間 B における演算へと変換可能である。

$$\iint f(x, y)dxdy = \iint f(x(u, v), y(u, v))|J(u, v)|dudv \tag{3}$$

$$= \iint g(u, v)|J(u, v)|dudv \tag{4}$$

$g(u, v) \equiv f(x(u, v), y(u, v))$ であり，$J(u, v)$ はヤコビアンである。分布関数も同様に空間 A から B へ写像される。

$$1.0 = \iint p(x, y)dxdy \tag{5}$$

$$= \iint p(x(u, v), y(u, v))|J(u, v)|dudv \tag{6}$$

$$= \iint q(u, v)|J(u, v)|dudv \tag{7}$$

$$= \iint P(u, v)dudv \qquad p() \text{ in A} \rightarrow P() \text{ in B} \tag{8}$$

$q(u, v) \equiv p(x(u, v), y(u, v))$，$P(u, v) \equiv q(u, v)|J(u, v)|$ であり，変数変換後にヤコビアンを掛けることで写像される。

以上の道具を用いて，空間 A と B の間に存在する不変量について考察する。空間 A における二つの分布，p_1 と p_2，を考える。これらを空間 B へ写像して得られる分布を P_1，P_2 とすると，当然 p_i と P_i の絶対的特性は異なる。p_1 と p_2 に対するバタチャリヤ距離は下記式で表記される。

$$BD(p_1, p_2) = -\ln \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy \tag{9}$$

これは，下記の様に空間 B における積分演算へ変換される。

$$BD(p_1, p_2) \tag{10}$$

$$= -\ln \iint \sqrt{p_1(x, y)p_2(x, y)}dxdy \tag{11}$$

$$= -\ln \iint \sqrt{q_1(u, v)q_2(u, v)}|J(u, v)|dudv \tag{12}$$

$$= -\ln \iint \sqrt{q_1(u, v)|J|}\sqrt{q_2(u, v)|J|}dudv \tag{13}$$

$$= -\ln \iint \sqrt{P_1(u, v)P_2(u, v)}dudv \tag{14}$$

$$= BD(P_1, P_2) \tag{15}$$

即ち，空間 A におけるバタチャリヤ距離は，空間 B における対応する二分布間のバタチャリヤ距離と等しくなる。この性質は，**式 (1)** の空間 A，B の対応付けに対して，強い制約を求めない。ヤコビアンによる変数変換が可能であれば，上記性質は満たされるため，一対一対応空間に対して付加的に要求される制約は，1)$x(u,v)$，$y(u,v)$ が偏微分可能で，導関数が連続，2) 空間 B の積分領域においてヤコビアン J が非零，のみとなる。結局，これらの条件を満たす，非線形変換を含む，広い変換群に対して，バタチャリヤ距離は不変となる。この変換不変性は，カルバックライブラ距離，ヘリンジャ距離などでも成立する一般的性質である。以上，各事象が分布として存在し，かつ，その推定が正確に行なわれれば，二分布間距離が非常に頑健な変換不変量として存在することを示した。この時，両空間の写像関数やヤコビアンを求める必要は無い。

2.2 不変事象間距離から普遍的に存在する不変構造へ

三辺の長さを規定すれば，三角形の形状は一意に定まる。同様に，ユークリッド空間に存在する n 点からなる幾何学構造は，${}_nC_2$ 個だけ存在する二点間距離を全て求めれば，(鏡像の曖昧性を無視すれば) その構造を一意に規定することになる。即ち，距離行列は幾何学構造を規定することになる。距離行列による構造定義は，タンパク質の構造解析など，広く用いられている方法である。距離行列と幾何学構造を等価であると考えるならば，空間に存在する N 個の分布群によって張られる距離行列，即ち，幾何学構造は，前節で数学的に導出した様に，一切変換不変となる。そして空間 A，B を二人の話者の音声音響空間 A，B とすれば，両者の間において不変構造が存在することになる。これはどの二話者でも成立するため，結局，話者非依存の普遍性を持つ（音響的普遍構造）。

この数学的性質は，非常に強力な枠組みとなると考えられる。従来筆者らは，英語学習者における英語母音群構造に対して，構造解析を行なってきた[12, 13]。話者／マイクなどの不可避的な音響歪みを除去し，外国語訛のみを構造歪みとして抽出することが目的であった。しかし，本来の構造不変性は，幅広い変換群で成立するため，学習者空間と教師空間との間に一対一対応があれば，外国語訛を超えて，構造の不変性・同一性を約束することになる。しかしこの場合，空間 A でのガウス性分布が空間 B では非ガウス性の分布として変換されるなど，分布形状の極端な歪みを生じることが予想される。例えば，変換後もガウス分布となるという制約の下で本数学的性質を使う等，積極的かつ妥当な制約導入によって，本性質は有効利用されると考える。逆に言えば，正確な分布の推定が可能であれば，それほど頑健な不変構造が，数学的には普遍的に存在する，ということである。本稿ではこの普遍的な不変構造の存在を基に，音声認知に関する種々の考察を行なう。

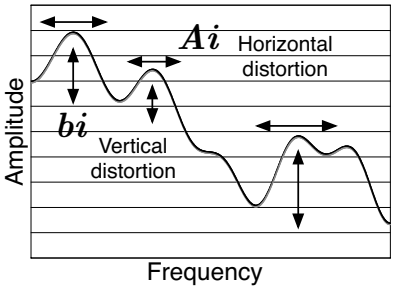

Figure 2: スペクトルの水平・垂直歪みと一次変換

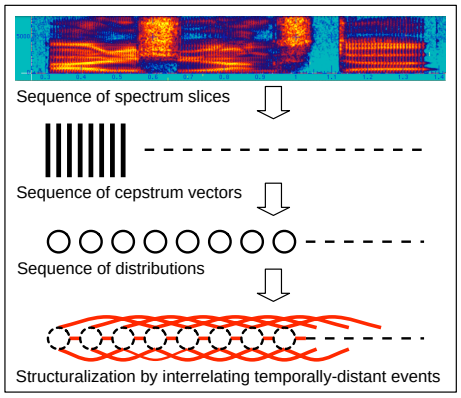

Figure 3: 事象間差のみを抽出して構成される不変構造

3 音声の構造的表象に対する実験的検討

従来より筆者らは，非言語的要因による音響変動をケプストラムの一次変換としてモデル化して議論してきた[8, 9]。これは，特にスペクトルの水平方向歪みが $c' = Ac$ として，垂直方向の歪みが $c' = c + b$ として記述できることによる（**図 2** 参照）。この場合，ガウス分布はガウス分布へと変換されることになる。これに対して**図 3** に示す様に，絶対項を全て捨象し，音事象間差異のみを求めることで，話者／マイク不変の構造的表象が得られる。実際に，孤立発声された 5 母音系列[2]をタスクとした音声認識では（語彙数 120 の孤立単語認識に相当），LPF などの前処理が必要ではあったが，一人の話者の音声で不特定話者音声認識が可能であることを示した[10, 11]。この実験では，4,130 人の話者の音声から構築された HMM よりも高い頑健性を示した。話者性を消去する，という方法論は，発音学習支援にも応用されている。特定の学習者と特定の教師の発音を，体格／性別／年齢といった違いを無視した形で，直接的・構造的に比較することが可能となっており，種々の興味深い実験結果が得られている[12, 13]。発音ポートフォーリオの提案，効率的学習のための教示生成，更には学習者分類などについて検討している。

図 3 に示す構造化による不変項の導出は「音声の非言語的特徴は時不変である」という仮定の上で成立する。即ち，話者性が時変であれば，不変項は導出されない。HMM 音声合成技術を用いて，時間的に話者性が変化する合成音

[2]連続発声を対象とした分布列推定方法がまだ確立できていないため，孤立発声母音系列という，人工的なタスクを用いた。

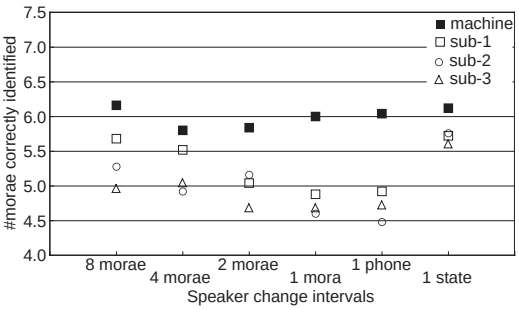

Figure 4: 非音声研究従事者を対象としたモーラ同定率

声を聴取させると，話者性変化頻度の向上によって無意味モーラ列音声のモーラ同定率が低下する様子が観測されている（**図 4**）[14]。なお，話者変化頻度を極端に上げた場合，HMM 音声合成の内部処理である時間方向のスペクトル平滑化によって話者性変化は消失されるとの予測が成立するが，実験結果も，その予測の妥当性を示している。

4　幼児は親の声の何を模倣しているのか？

幼児は親の声の模倣を通して音声言語を獲得すると言われている（音声模倣）[6]。しかし幼児は，親の声そのものを模倣しようとはしていない。太くて低い声を出そうと努力している幼児はいない。音韻意識が未熟である彼らは，個々のモーラ（話者非依存の音シンボル）を一つずつ生成する，という術は少なくとも意識的には不可能である。となると彼らは親の声の何を模倣しているのだろうか？九官鳥による音声模倣では，話者性までも真似ることが知られている[15]。優秀な九官鳥は，その音声模倣を聞いただけで飼い主が分かるが，どんなに優秀な幼児の音声模倣を聞いても，飼い主（親）の同定は不可能である。音響的な音声模倣と言語的な音声模倣は何が違うのだろうか？

「幼児の音声言語獲得は，分節音の獲得の前に語全体の音形・語ゲシュタルトの獲得から始まる」との主張が正しければ[3]，この語ゲシュタルトの音響的実体には，非言語的情報は含まれないはずである。もし含まれていれば，幼児は父親の声が出せるよう，日々努力するはずである。筆者らはこの語ゲシュタルトの音響的定義について，多くの発達心理学・言語獲得研究者に問いかけてみたが[16]，残念なことに，明確な答えは得られなかった。

近年の脳科学の進歩により，聴覚音声学の議論は，蝸牛から，聴覚皮質のモデリングに移行しつつある[17]。脳科学における多くの知見は偶然によって齎されている[18, 19]。交通事故や，一部の医師の不適切な処置が原因で不幸にも脳損傷を負った患者を通して多くの知見が得られている。動物実験でも同様，偶然的な刺激提示によって重要な知見が得られている。前頭葉，海馬，扁桃体の機能，更には，ミラーニューロンなどは良い例である。脳は研究者の机上の議論を超えた処理を行なっている，と解釈することもできる[19]。さて，聴覚皮質モデリングであるが，視覚皮質のような定説が存在する状況には無いが，幾つか興味深い主張がある。まず「音声の言語的情報と非言語的情報（話者の情報）は分離されて処理されている」との主張である[17, 20, 21]。特に [21] では，音楽と音声とを対比し，音声の言語情報は，音声の動きの情報（speech motions）によって伝搬されると主張している。音楽で言えばメロディーである。一方「話者の同定は音楽で言う楽器の同定に相当し，それは時不変の情報として処理される」と主張している。**図 3** に示した音声の構造的表象は，音声を「音の運動」と考え，その運動（コントラスト）成分のみを抽出する形となっている。即ち音声から「音であること」を一切捨て去った物理表象である。何が動いているのかは不明である。「動きだけを抽出した時に，話者／年齢／性別を超えた頑健な不変表現が数学的に入手できる。それこそ言語である」と主張するのが音響的普遍構造である。

近代言語学の祖ソシュールが一世紀以上も前に興味深い主張をしている[22]。The important thing in the word is not the sound alone but the phonic differences that make it possible to distinguish this word from all others. 即ち，音ではなく，音的差異の重要性を説いている。差異を捉えることで単語が同定できる，との主張である。彼はまた Language is a system of conceptual differences and phonic differences. と主張している。「言語＝差異・動きのシステム」である。分布間差異を集めたものが頑健な不変構造を成し，それを用いた語同定が可能である。この不変構造こそ語ゲシュタルトではないだろうか。

音声が伝搬する情報は，言語／パラ言語／非言語情報と分類される。各情報を担う音響量に着眼すると，言語及び非言語情報は声道情報となるため，スペクトル包絡に相当し，パラ言語情報は音源情報となるため，F_0，パワー，継続長に相当する。即ちソース・フィルタの分離である。幼児の音声模倣は，親の声からまず非言語情報を分離すると考えられる。音声＝［言語＋パラ言語］＋［非言語］という枠組みである。しかし，音声科学・工学が構築した枠組みは，音声＝［言語＋非言語］＋［パラ言語］という枠組みである。調音音声学の価値観に基づけば，声道と音源を分離する，自然かつ妥当な枠組みである。しかし，音声コミュニケーションの観点から考えると，この枠組みでは幼児の音声模倣問題は解けない。非言語情報を頑健に分離する術が無いからである。人と音声との遭遇は聴取であって，生成ではない。しかし，科学は音声と生成を通して遭遇した，というのは歴史的事実である[23]。音声科学が実験科学である以上，それは時代の技術的制約の下で議論を重ねなければならない。聴覚音声学は観測技術の未熟さから，調音・音響音声学と比較して，その進展

[3]なお，日本人乳児が [r] と [l] を弁別できることが広く知られているが，これは 2 音の弁別ができるのであって，[r] を/r/として同定している訳では無い。同定能力の獲得の前に，まず，弁別・区別，即ち差異の知覚が可能になることは重要である。

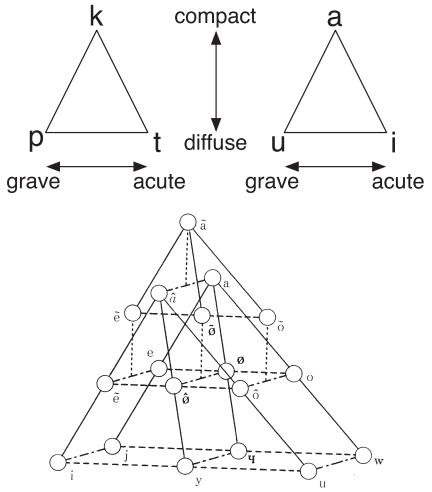

Figure 5: 母音・子音三角形[24] と仏語母音群構造[25]

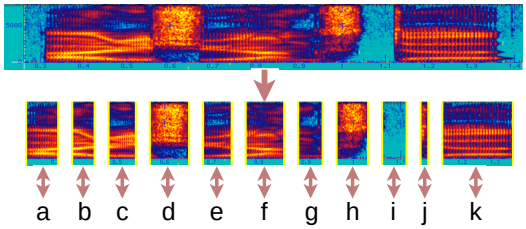

Figure 6: 音声ストリームの細分化と絶対音感的要素同定

が遅れざるを得なかった。脳科学の進展によって「話されたもの」としてではなく「耳に届くもの」として音声に焦点が当たった時に，従来の枠組みでは想像できない情報処理を脳が行なっていたとしても何ら不思議ではない。

ソシュールの phonic differences という言葉はやがて，ヤコブソンの弁別素性，即ち構造音韻論へと引き継がれる。音そのものではなく，音/x/と音/y/はどう違うのか，その違いを定性的に表現するために弁別素性が使われた（**図 5** 参照）。つまり，音素群が成す外部構造の議論である[25]。やがて，弁別素性はそれが束となって音素を表象する，即ち音素の内部構造の議論に使われるようになる[24]。この素性の束としての音素は音楽の「和音」をメタファーとして生まれた[26]。和音＝音素，音符＝素性，である。音楽学では，音楽には横の構造（メロディー構造）と縦の構造（ハーモニー構造）があると説く。和音は縦の構造であり，ヤコブソンの素性による内部構造の議論は音楽の縦構造を発端としている。一方，筆者らが提唱する音声の構造的表象は当然のことながら，音楽の横構造に相当する。縦と横の構造，どちらが音楽にとってより本質的かと言えば，当然，横構造である。和音の無い音楽はあれど，メロディーの無い音楽は存在しない。

筆者らの知る限り，音声工学において横構造の議論が皆無であり，縦構造の議論[27, 28, 29] が多い理由は，音声の表記方法に起因すると考える。音声を（話者非依存の）音シンボル列として表記し，音声をシンボルに対応させて区切れば，その時点で横構造は消失する。例えば Bloomfield は，

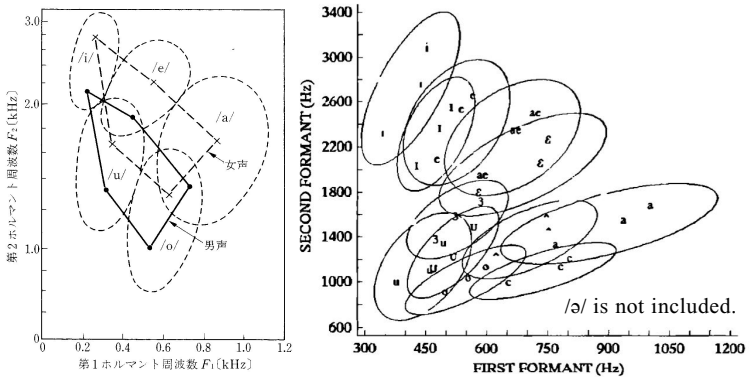

Figure 7: 日本語 5 母音と米語 12 母音の F_1/F_2 図[31, 32]

Writing is not language, but merely a way of recording language by visible marks. と述べているが，文字言語は本来音声言語の副産物であり，5 万年以上の音声活動を通して人類が造り出した「音声言語の視覚化技術」でしかない[4]。即ち「結果」であって「原因」ではない。しかし「結果」を「前提・原因」として捉え，**図 6** に示す様に，音声の物理現象を切り刻んで来た，という歴史は否めない。この一次元的音声視覚化技術は物理的に正しいのだろうか？そもそも科学は要素還元主義の上に構築され，その枠組みの限界が指摘されたのはごく最近である[30]。

5　音韻の意識は音声言語運用に必要なのか？

提案している音声表象は音声ストリームをメロディとして捉え，その横構造を頑健な不変項として導出している。孤立発声母音の系列という人工的タスクではあるが，個々の音事象の絶対的物理特性は一切用いずに，個々の音事象の同定は一切行なわずに，単語の同定が可能であることを示した[10]。音楽の場合，横構造を通して各音事象の機能的・相対的価値を感覚し，「ドレミ」が聞こえてくる[5]。音声の音韻知覚も同様の枠組みとして捉えられないだろうか？全体を通して要素同定する，音声の相対音感である。

音楽の場合，如何なる鍵盤（即ち物理特性）も「ド」になれる。逆にある鍵盤は「ドレミ…」のいずれにもなれる。しかし音声の場合，例えば F_1/F_2 図において任意の点の音を「あ」と知覚できるか，と考えれば，それは困難である。**図 7** に示す様に日本語の場合，男女を考慮しても 5 母音は凡そ分離している。音韻とその物理実体との対応が凡そ一対一，即ち，絶対音感的である。しかしフォルマント周波数が発声者の声道長に依存していることを考えれば，例えば目玉親父やウルトラマンが日本語母音を発声したとすると，これらの分布群は大きな重なりを呈することになる。一対一対応の崩壊である。このような場合，音の絶対量に基盤を置く処理系は機能せず，音と音の相対量に基盤を置いてこそ，頑健な処理系が期待できる。

[4] 文字起源は象形文字であるため，本来文字は意味の視覚化技術であり，音の視覚化技術ではない。表音文字は単なる借り物技術でしかない。
[5] 人が沢山いるそうである，としか第一著者は言えない。

アニメの世界では音声は相対音感的でなければならない。

アニメの世界を想像しなくても，相対音感の世界を創成することは容易である。母音の数を増やせばよい。**図 7** には米語 12 母音の F_1/F_2 図についても示している。成人男性・女性・子供（10〜12 歳）139 名の/h V d/から得られた結果である[32]。なお，音質が容易に変動する/ə/はこの図には含まれていない。これだけの重なりは，複数話者のデータを同時に表示するから生じるのであり，話者別に示せば当然重ならない。この事実を顧みずに，母音毎に，複数話者データに対して物理的な絶対量を統計的にモデル化しても，母音認識は困難となる。日本語は絶対音感的，米語は相対音感的なのだろうか？何れの場合も，個々の音はシンボル化される。音楽の場合でもドレミは階名としても（移動ド），音名としても（固定ド）使われている。

相対音感者の多くは，言語化できない相対音感者である。メロディーの記述を，音名／階名で行なうのではなく，「ラ〜ラ」即ちハミングで行なう相対音感者である。彼らも主音は認知しており[3]，音楽の横構造を認識しているが，主音に対して「ド」を対応させることが困難である。そもそも音高（基本周波数）と「ドレミ」という声（スペクトル包絡）とは無関係であり，これを恣意的に結びつけたのが階名である[6]。音声に対して「言語化できない相対音感者」とはどのような存在になるのだろうか？他者が歌った歌を「ラ〜ラ」として再生する際に頻繁に移調されることを考えれば，ある話者の発声を移調して再生することは，「繰り返し発声」に相当する。一方，曲を「ドレミ」に落とす作業はどうなるであろうか？スペクトル特性とは全く関係の無い，「声」に対して恣意的に関連付けられた「モノ」を考えれば明らかなように，それは「(表音）文字」に落とす作業となる。以上の考察から得られる帰結は「相対音感的な音認知が不可避的に要求される言語の場合，文字の読み書きに困難を覚える人が多い」となるが，こんな考察，意味があるのだろうか？

第一著者は，このような無意味かもしれない考察の最中に失読症（dyslxeia）を知った。「頭が良いのに，何故か本が読めない」方々である[33, 34, 35]。具体的な症状は様々であるが，共通項として存在する症状が音韻意識が希薄，即ち，単語音声に対して，それを個々の音に分割したり，個々の音が連結して単語音声になる，ということを感覚することが困難な方々である。**図 6** の枠組みを理解することが困難な方々である。幼児の音声認知をそのまま引きずっており[33]，個々の音をカテゴリとして知覚するのが困難である一方で，異音の区別は健常者よりも成績が良い[35]。これは [r] を/r/というカテゴリとして同定できないが，[r] と [l] が区別できることに相当する。米国では程度の差こそあれ，約 20%の人が失読症である[33]。政治家，作家，起業家，学者にも失読症者はおり，グラハム・ベルもその一人である。彼に音声認識・合成器を作らせても，音シンボル列と音声間の変換技術など作ら（れ）なかったはずである。「そんなモノの上に言語は出来ていない」と主張したであろう。幼児の言語獲得は，彼らの認知能力の未熟さが，個々の音韻を意識させないのだろうか？無意識下では音韻を操作しているのだろうか？或いは，個々の音韻意識は音声言語運用に不要なのだろうか？**図 6** の「音声 ↔ 音シンボル列変換」に難を示す多数の音声ユーザの存在を，音声科学・工学者はどう考えるべきなのだろうか？

言語化できる相対音感者が時として犯す勘違いとして，次のようなものがある。全ての長調の曲が「ハ長調」として聞こえる，というものである。凡そ全ての曲は主音で終了する。即ち，階名で「ドレミ」が聞こえて来る相対音感者は，曲の終わりは全て「ドー」と聞こえる[7]。常に「ドー」と聞こえるから，その箇所では同一の鍵盤を押している，即ち，同一の物理音が出ている，と解釈した訳である。機能的・相対的等価性が物理的等価性を上書きし，異なる音群に対して同一物理音を認知させた訳である。そのような相対音感者に対しては，絶対音感者が「それは物理的には錯覚，勘違いの一種である」と説明する。機能的・相対的等価性を物理的等価性と入れ違えた結末であると説明する。音声科学・工学では，話者 A の音声中のある音が音韻「あ」と感覚され，話者 B のある音も音韻「あ」と感覚された場合，**図 6** に示した音声ストリームの細分化を行ない，両話者の該当区間の物理現象に何らかの絶対的同一性を期待する。その二音の物理的相違は明らかであるにも拘らず，数千，数万人の話者から「あ」と感覚される音声区間を集め，統計的にモデル化する。音韻とは心的表象である[36]。心的表象とは物理実体が存在しないことを意味する。よってその心的表象は，物理的にはある種の「錯覚，勘違い」の産物ということになる[37]。音響音声学，音声工学が大前提とする**図 6** の枠組みは，物理的に妥当なのだろうか？音楽の相対音感者の勘違いは音楽の絶対音感者が是正してくれた様に，**図 6** の枠組みが研究者の単なる勘違いであるとするならば，音声の絶対音感者が，彼らを是正してくれるのだろうか？

6　究極の音声絶対音感者と音声言語

極端な絶対音感を持つ奏者は，オーケストラ／ホールが変わる度に十分な耳慣らしが必要となる。基準音がオーケストラ／ホールによって，数 Hz 異なるからである。参照パターンとして絶対項を持ってしまうと（例えば，基準音＝ 440Hz），環境の変化に対して柔軟に対応できなくなる。音声の極端な絶対音感者は話者 A の「おはよう」と話者 B の「おはよう」の同一性の認知が困難になると考えられるが，自閉症者の一部に，特定話者の音声のみ言語メッセージになる者がいる[1]。自閉症は端的に「関係の

[6]「ドレミ」という命名は僧侶の名前の第一音節から来ている。

[7]らしい。くどいようであるが，第一著者には皆目見当がつかない。

病」と言われるように[38]，入力される情報の整理整頓が困難であり，個々の要素的事象を丹念に記憶する。日付，曜日，電話番号，住所など互いに無関係なものを膨大に記憶する一方で，物事の因果関係や複数の刺激群が成すパターンの抽出，事象の抽象化に困難を示す。そのため，目の錯覚などが起き難い。マガーク効果が起き難い。顔の要素的特徴を覚える一方で，顔を見て表情や話者を同定することが苦手である。優れた音感を持ち，絶対音感者が多い。一言で言えば，ゲシュタルト知覚が困難である[39]。

第一著者にとって「ドレミ」とは音名であるため「曲が階名として聞こえる」という事実は想像を絶する。「ソ」が「ド」と聞こえる，というのは「え」が「あ」と聞こえる，というのに等しい。勘違いか錯覚の類いではないか，とさえ考えることもある[8]。極端な音声の絶対音感を持つと考えられる自閉症者にとって，物理的に異なる特性を持つ話者 A の音と話者 B の音を「同一音」として認知する健常者の感覚こそ，想像を絶するものであると推測する。彼らが「勘違いか錯覚の類いではないか」と主張しても不思議ではない。異なる二音を「あ」と感覚できる健常者の認知能力が，音の絶対項に基づくものなのか，あるいは，音間の相対項に基づくものなのか，彼らこそ，その回答をもたらしてくれるもの，と期待されるが，残念なことに，彼らの多くは口を開かない。何故なら，極端な絶対音感を持つ自閉症者は，音声言語を持たないからである。二話者の「おはよう」の同一性が認知できなければ，音声言語が破綻するのは自明である。音声言語は，ある種の錯覚・勘違いの上に成立する，と考察することもできる。音声言語を持たない自閉症者の中には，ごく稀に，文字言語を通して言語コミュニケーションを開始する場合がある[1, 40]。音は全て聞こえているにも拘らず，聞こえ過ぎるが故に，文字（視覚図形）言語が第一言語となる。自閉症者は，常に変化する環境を頑健に対処する術を持ち合わせていないと言われる。文字は変わらない。しかし，音声はいつも変わる。だから図形言語が第一言語となる。確かに，人，場所，時，あらゆる要因が音声の絶対項を変える。しかし着目する時間長において，その要因が時不変であれば，構造は一切不変である。変わることが許されない。

音響空間を [音素数]^3 の部分空間に分け，各々の独立性を仮定して各空間における観測量を絶対的にモデル化し，保持するのが現在の音響モデリング技術の常套手段である triphone である。その結果，環境が変わる度に耳慣らし（音響モデル適応／特徴量正規化）が不可欠となる。似ていないだろうか？筆者らには，音声認識における音響モデリングは，自閉症者の音感そのものであるように思える。問題の本質は，**図 6** に示した「音声 $\leftrightarrow$ 音シンボル列変換」を物理的前提として音声の物理現象を解析することにあると考える。音声ストリームに対して，聴取者が感覚する音韻列を並べ，各音韻に対応する音声区間を切り出す。音韻は話者不変であるが，一方の物理現象は，人，場所，時，あらゆる要因がこれを変え，多様性問題に直面することになる。従来の音声科学・工学は「集めること」でこの問題を回避しようとしたが，本稿は，これを直接的に解く方法を提供している。筆者らは，**図 6** に示す一次元的音声視覚化技術は，物理的には，バグのある技術であると主張する。このバグのために「音声言語の正規ユーザ」が悩んでいても，何ら不思議ではない（失読症）。このバグのために，音シンボルの物理的対応物の不変性を信じて，その物理的対応物を絶対的に記憶する方々が音声言語運用に悩んでいても，何ら不思議ではない（自閉症）。

工学システムと自閉症者との類似性の議論は，古くはロボット工学に見られる。フレーム問題に端を発してロボットと自閉症児との類似性が議論されており，現在でも続いている[41, 42]。自閉症者は環境の些細な変化に非常に弱い側面を見せる。花瓶の位置が変わっただけでパニックに陥る場合もある。同様に，指定された部屋の情報を全てインプットされたロボットが，猫の来訪など，予期せぬ出来事にパニックに陥る。多様に変化する環境を頑健に対処できない両者に，工学者が，自閉症セラピストが，互いの類似性を認め合った経緯を持つ。環境の多様性を生き抜く術を与えるべく，工学者・セラピストが協力している。

言語発達に遅れの無い自閉症をアスペルガー症候群と言うが，彼らの音声言語活動は，やはり健常者とは異なる側面を示す[43, 44]。音声をまず文字化し，テキストを通して理解しようとする。そのため言語の論理面（文字面）だけの解釈となり，パラ言語的情報など文字化で消失する情報の処理が困難である。その音声が発せられた場・文脈を通して発言を解釈しようとせず，表層文だけに基づいて解釈を試みるため，場に合った対応ができず，多義性を解決する，行間・真意を読むなどの処理が苦手である。元々音声は苦手であり，電話音声などは特に困難である[43]。これらは，現状の音声対話システムに対しても，広く当てはまる性質である。アスペルガー症候群を患う者を家族に有する者は「計算機に音声コマンド入力するようなもの」と，彼らとの音声対話を記述している[43]。彼らの多くは自らを「地球生まれの異星人」と呼ぶ。感覚系・知覚系が健常者とは大きく異なるからである。「音声認識技術は，人間シミュレータを目指す必要は無い」という議論は古くからある。システムの入出力さえ模擬できれば，内部処理の実装まで模擬する必要は無い，という議論である。しかし，実際に構築したシステムは，そのビヘービアのみならず，内部処理の実装に至るまで非常に類似している「現実の対象物」が存在している。残念ながらそれは人間ではなく，自称異星人である，というのが筆者らの意見である。ヒューマノイドという名称で呼ばれる機械が巷に溢れているが，この「異星人」の存在を知る筆者らには，

[8] 実際には「ド」の意味が異なるので，勘違いでも錯覚でも無い。

少なくともその機械の音声処理系に関して「ヒューマン」という名称を使うことに強い抵抗を感じざるを得ない[9]。ロボット工学同様，音声の多様性を生き抜く術を両者に与えるべく，議論を重ねる必要があると考える。

7　生物進化と音の情報処理　〜絶対と相対〜

多くの動物は刺激間の相対的特性よりも，対象とする刺激の絶対的特性に基づいた処理を行なう傾向にあることが知られている。これは，相対的特性に基づく処理系の方が，より高度な認知能力を要求するからである，と考えられている[46]。音高に関しては，ラットやオオカミは絶対音感であることが報告されている。アカゲザルも基本的には絶対音感であるが，絶対性に基づく処理が失敗すると，相対性に基づく判断も行なう[47]。またニホンザルも同様，絶対音感としての処理が基本となっており，局所的な手がかりに着目する様子が報告されている[48]。このように生物進化の過程の中で音高処理が，絶対的な属性から相対的な属性へと遷移してきた様子が論じられている[49]。

本稿で論じてきた音声の構造的表象は，音高ではなく，スペクトル包絡という形で物理的に観測される音質に対する相対的な処理を対象としている。この音質に関する相対処理というのは，ヒト以外の動物では考察が困難である。そもそも，ヒトがこれだけ多様な母音を生成できるのは，二足歩行による喉頭の下落により，口腔に十分な空間を有するようになったからである。調音器官を制御して口腔を変形させることで，様々な共鳴パターンを生じさせ，これが様々な母音の生成を可能とした。当然口腔のサイズ／形状は話者依存であるため，音質の多様性は拡大する。本稿では，口腔のサイズ／形状に起因する静的な音響歪みを頑健に消失させる方法論として，音声の構造的表象を提案し，様々な観点から本手法を考察した。

8　まとめ

筆者らが提唱する音響的普遍構造が頑健な変換不変性を有することを数学的に示し，相対音感としての音声認知を通して，言語獲得，失読症，自閉症を考察した。本表象では音声の多様性問題が何ら問題になり得ず，また，パラドックスとも言われる，失読症や自閉症の音声認知についても，凡そ自然な考察で説明可能であることを示した。しかし，本考察がこれら障害の全容を網羅している訳ではなく，例えば失読症と自閉症の合併症が存在するのも事実である[50]。また，幾つかの凡そ典型的と考えられる症状について示したが，これらの障害は非常に多様な症状を呈しており，記述した各項目が常に観測される訳では無いことを断っておく。しかし，自閉症と音声認識技術の類似性について考察したように，自らを異星人を呼ぶ彼らのビヘービア及び認知特性は，より人間らしい機械を構築することを目指す工学者にとって，非常に有益な情報を提供していると筆者らは考える。**図 6** に示す音声の分節化及び要素の絶対的同定は，問題の要素還元に基づく方法論である。要素間の独立性を仮定した方法論である。その仮定に本質的な不備がある場合，要素分割とは異なる枠組みが必要となる。本稿はその一提案である。

[9]改名すべきモジュールが音声処理系だけであるかどうかは，言及しない。しかし，アスペルガー症候群の方々の身体の運動制御が，健常者のそれとは，やはり異なっていることを指摘しておく[45]。

参考文献

[1] 東田他, この地球にすんでいる僕の仲間たちへ, エスコアール出版社 (2005)
[2] 奥平, 自閉症の息子ダダくん 11 の不思議, 小学館 (2006)
[3] 谷口, 音は心の中で音楽になる, 北大路書房 (2003)
[4] 東川, 読譜力－「移動ド」教育システムに学ぶ, 春秋社 (2005)
[5] 加藤, コミュニケーション障害学, 20, 2, pp.84–85 (2003)
[6] 早川, 月刊言語, 35, 9, pp.62–67 (2006)
[7] 原, コミュニケーション障害学, 20, 2, pp.98–102 (2003)
[8] 峯松他, 信学技報, SP2005-12, pp.1-8 (2005)
[9] 峯松他, 信学技報, SP2005-131, pp.121-126 (2005)
[10] 村上他, 信学技報, SP2005-14, pp.13-18 (2005)
[11] 村上他, 信学技報, SP2005-130, pp.115-120 (2005)
[12] 朝川他, 信学技報, SP2005-24, pp.25-30 (2005)
[13] 朝川他, 信学技報, SP2005-156, pp.37-42 (2006)
[14] 峯松他, 信学技報, SP2004-27, pp.47-52 (2004)
[15] 宮本, 音を作る・音を見る, 森北出版 (1995)
[16] N. Minematsu, *et al.*, "Universal and invariant representation of speech," Proc. Int. Conf. Infant Study (2006)
[17] 柏野, 月刊言語, 33, 9, pp.102–107 (2004)
[18] M. Spitzer, 脳・回路網の中の精神, 新曜社 (2001)
[19] 茂木, 心を生みだす脳のシステム, 日本放送出版協会 (2001)
[20] K. S. Scott *et al.*, Trends in Neurosci., 26, pp,100–107 (2003)
[21] P. Belin *et al.*, Nature Neurosci., 3, 10, pp.965–966 (2000)
[22] F. D. Saussure, Course in general linguistics, McGraw-Hilll Humanities/Social Sciences/Langua (1965)
[23] 前川, 音声研究, 8, 3, pp.35–40 (2004)
[24] R. Jakobson *et al.*, Preliminaries to speech analysis, MIT Press, Cambridge, MA (1952)
[25] R. Jakobson *et al.*, Notes on the French phonemic pattern, Hunter, N.Y. (1949)
[26] S. E. Blache, The acquisition of distincitve features, Univ. Park Press (1978)
[27] K. N. Stevens, J. Phonetics, 17, p.3–45 (1989)
[28] L. Deng *et al.*, Speech Comm., 33, 2–3, pp.93–111 (1997)
[29] M. Ostendorf, Proc. ASRU, pp.79–84 (1999)
[30] M. M. Waldrop, 複雑系, 新潮社 (2000)
[31] R. K. Potter *et al.*, JASA, 22, 6, pp.807–820 (1950)
[32] J. Hillenbrand *et al.*, JASA, 97, 5, pp.3099–3111 (1995)
[33] 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)
[34] 石井, 科学技術政策研究所・科学技術動向 45, pp.13–24 (2004)
[35] W. Serniclaes *et al.*, Cognition, 98, pp.B35–B44 (2005)
[36] H. A. Gleason, An introduction of descriptive linguistics, Holt, Rinehart & Winston (1961)
[37] A. J. Lotto *et al.*, Chicago University Society, 35, pp.191–204 (2000)
[38] 酒木, 自閉症の子どもたち, PHP 研究所 (2001)
[39] U. Frith, 自閉症の謎を解き明かす, 東京書籍 (1991)
[40] R. Martin, 自閉症児イアンの物語, 草思社 (2001)
[41] 渡部, 鉄腕アトムと晋平君, ミネルヴァ書房 (1998)
[42] J. Nade, "The developing child with autism," Tutorial Session of IEEE Int. Conf. Development and Learning (2005)
[43] 泉, 僕の妻はエイリアン, 新潮社 (2005)
[44] 榊原, アスペルガー症候群と学習障害, 講談社 (2002)
[45] ニキリンコ, 自閉っ子, こういう風にできてます！, 花風社 (2004)
[46] D. J. Levitin *et al.*, Trends in Cognitive Sciences, 9, 1, pp.26–33 (2005)
[47] A. A. Wright *et al.*, Journal of Experimental Psychology, General, 129, pp.291–307 (2000)
[48] A. Izumi, Journal of Comparative Psychology, 115, pp.127–131 (2001)
[49] M. D. Hauser *et al.*, Nature Neurosciences, 6, pp.663–668 (2003)
[50] 月文他, 自閉症者からの紹介状, 明石書店 (2006)