音声の構造的表象と音声の相対音感*

○峯松信明（東大新領域），西村多寿子（東大医学系），櫻庭京子（清瀬市障害福祉センタ）

1　はじめに

幼児が親の発声を “真似て” 言葉を覚える。日常的な微笑ましいシーンである。しかしこのシーンに対して「親の声の個人性までを真似ようとしてませんが，（音韻意識の無い，音声を音韻に分割できない）彼等は親の声の何を真似ているのですか？」と問いかけた時，この問いに対する物理的に妥当な回答を音声科学・工学はまだ用意できていない。こんなシーンですら，その物理的過程を記述できていないのである。次に父親が鼻歌を歌い，それを幼児が真似たとする。「二人の鼻歌の基本周波数は異なりますが，今度は何を真似ているのでしょう？」と尋ねると，「相対的なパターンでしょう」と誰もが答える。本稿は，前者の問いについても同様の回答が可能であり，それが言語学，言語障害学，神経生理学などの観点から見て妥当であることを示す。と同時に，従来の音声科学・工学が構築した枠組みが持つ「不備」を指摘する。

2　音声の構造的表象　～音響的普遍構造～

2.1　音声＝言語＋パラ言語＋非言語

周知のように，音声には「言語，パラ言語，非言語情報」が含まれる。この観点に立てば，上記の問いの答えは「言語＋パラ言語情報」を真似ている，となる。つまり「非言語情報のみを分離して」真似ている，ということである。しかし「音声から非言語情報を物理的に分離するとは？」と問い掛けた時に議論は暗礁に乗り上がる。何故なら，音声科学・工学は非言語情報の分離を「非言語情報で和をとること（数千～数十万人）」で，物理的に実装してきたからである。

$$g(\text{言語情報}) = \sum_{\text{非言語情報}} f(\text{言語情報，非言語情報})$$

視覚生理学では「第一次視覚野以降，色，形，キメ，動きを処理する部位が各々存在し，その結果が連合野でまとめられる」とする情報処理モデルが確立している。これに対して聴覚生理学でも最近になって漸く，第一次聴覚野以降の処理モデルを呈するようになった [1, 2]。そこでは「言語情報と話者情報は異なる部位で処理されている」ことが示されている。（ほぼ）母親，父親の声だけで，何故幼児は誰の声でも楽々と認識できるようになるのか？もし話者性が物理的に分離できるならば，本来この問題は問う価値すら無い。SAT の枠組みは「音声認識システムが持つ音響モデルの話者性は，ある固定話者でよい」ことを示す（常時，適応をかける）。しかし，筆者らは常時適応をかけて音響スコアを求める場合，そのスコアは入力音声から音的差異（コントラスト）を抽出すれば，明示的な適応処理無しに求まることを数学的に示した [3]。

2.2　音響的普遍構造 [3, 4]

音声に不可避的に混入する歪みとして，線形変換性の歪み（声道長，聴覚特性の違い），畳み込み歪み

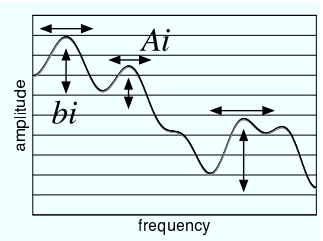

Fig. 1　不可避的な歪み

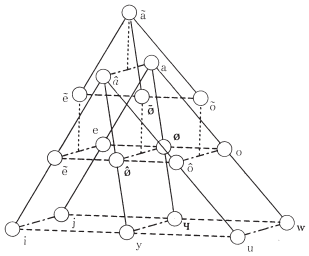

Fig. 2　構造的音韻論

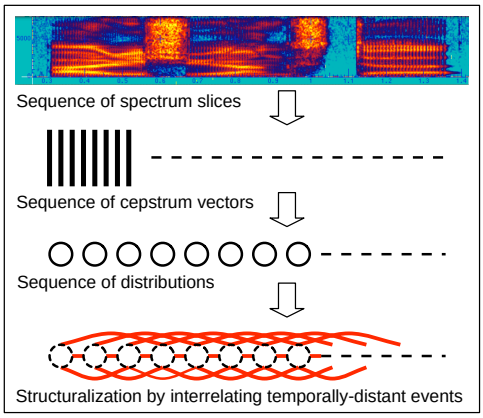

Fig. 3　発声の構造化

（マイク特性，話者性の一部）を考える。加算性の背景雑音は考えない。上記二種類の音響歪みはケプストラムの一次変換 $c' = Ac + b$ で近似される。**図 1** に示すように，線形変換性歪み，畳み込み歪みはスペクトルの水平方向（A），垂直方向（b）の変動に対応する。**図 3** に示している，ケプストラム系列を分布（ガウス混合分布）系列に変換した後に計算される分布間距離（バタチャリヤ距離等）群，即ち，距離行列は一次変換不変な物理量である。n 点から成る幾何学構造としての n 角形は，全ての 2 点間距離，即ち $n \times n$ の距離行列によってその構造は一意に規定される。同様に分布間距離群として求められた距離行列も（ある非ユークリッド空間における）幾何学的構造を規定する。**第 2.1 節**に示したように，この距離行列のみから，話者適応後の音響照合スコアが近似的に求まるため，（原始的なタスクであるが）孤立母音系列（系列長 5，種類数 120）の音声認識を行なったところ，認識率は 100%となった。即ち，スペクトル包絡やフォルマント周波数など，音声の絶対的特性を用いることなく，音声が認識可能であることが示された [5]。

3　言語学と音響的普遍構造

この音声の構造的表象はどのような意味を持つのだろうか？言語学的意味を考える。「単語（音声）で大切なのは音そのものではなく，音的差異であり，その差異によって単語は他の単語から識別出来る」と述べたのは，近代言語学の祖，ソシュールである。彼は「言語の要素は絶対的には定義できず，他の要素との

*Structural representation of speech and relative sense of speech sounds.
Nobuaki Minematsu, Tazuko Nishimura, and Kyoko Sakuraba

関係によってのみ定義可能となる」と述べている。即ち，上記の音声認識実験は「約 100 年前の言語学者の主張が数学的にも実験的にも正しかったことを示した」実験との位置づけを持つ。周知のようにソシュールの発言はその後，構造主義と呼ばれる思想を生むが，構造主義に基づいた言語音論，即ち，構造的音韻論がヤコブソンによって展開された。**図 2** は，ヤコブソンによる母音群の幾何学構造である。音響的普遍構造はこの数学的，物理学的実装に他ならない。

4 言語障害学と音響的普遍構造

提案した音声認識手法は，単語を音韻の系列として捉えずに，単語を全体的にかつ構造的にのみ，表象している。これは単語の識別において個々の音韻を同定することは不要であり，個々の音韻モデルは不要であることを示唆する。言い換えれば「音声コミュニケーションは問題無く行なわれるが，日本語音声を平仮名列として書き起こせない，或は，平仮名系列を読み上げることができない」症状を呈する日本人話者の存在を示唆するが，非常に類似した症状を呈するのが音韻性失読症（難読症，dyslexia）である。

「知的な遅れや視聴覚障害がなく，十分な教育歴と本人の努力にも拘らず，その知的能力から期待される読字能力が低い」と定義されている [6]。日本では症例報告が少ないが，欧米では軽症も含めれば 15%の人が失読症であると言われており（視覚性失読症も含む），広く知られた障害である。文化人，知識人，政治家などにも見られる症例である[1]。当初視覚障害としての可能性が追求されたが，現在では，多くの研究者が音韻意識の欠如を原因として考えている。即ち，音声を個々の音の線状結合として捉え，それを音に分割することができる能力は，音声コミュニケーションの必要条件ではない。その能力は「書き起こす」時にのみ必要な能力と言える。同様のことが音楽の絶対音感（採譜能力）にも言える。多くの音楽家が「絶対音感は採譜の時にあれば便利な能力であるが，音楽活動の必要条件ではない」と述べている。失読症者自身の記述によれば，彼等は「物事（事象群）の全体的な様態の知覚がまず起き，その一方で，個々の事象に注意を向けることが困難である」[7] とのことである。

「個々の事象の個別的な知覚がまず起き，その一方で，事象群の全体的な様態に注意を向けることが困難である」[8] 症状を呈するのが自閉症である。先天性の脳機能障害であり，社会的相互交渉（対人関係）の障害，言語コミュニケーション（特に音声）の障害，強いこだわり（常同性の選好），想像力の障害などで特徴付けられる。一言で「関係の病」と述べている研究者もいる。絶対音感者が多い，マガーク効果などの錯覚が起こり難いなど，現象を要素分割して知覚する様子が報告されている。また，電話番号，住所録，時刻表など，互いに関係の無い要素群を丸暗記する記憶力を持つ場合もある。彼等は示されたパターンを「そのまま」記憶するのは得意であるが，それに変形が施されて環境の中に存在した場合，それを発見するのが苦手である。即ち要素の常同性に固執し，環境の些細な変化に対してパニックを起こすなど，非常に弱い面を示す。この環境変化に対する弱さは人工知能のロボット研究が直面したフレーム問題と等価であり，自閉症児とロボットとの類似性が議論されている。筆者らは音響空間を領域分割し，各領域を互いに独立にモデル化し，記憶する（要素還元主義に基づく）従来の音響モデリングも自閉的であると指摘した [3]。興味深いことに，文字による言語コミュニケーションはとれるが，音声は母親の声のみ意味へと変換できる自閉症児がいる。母親以外の声は「音としては聞き取れるが，意味へ変換出来ない」そうである [9]。本当の意味での音声の絶対音感者であると筆者らは考える。

[1] アルバート・アインシュタイン，ジョン・レノン，トム・クルーズらが失読症であったことは知られた事実である。自らの名前が研究所に冠されたグラハム・ベルもその一人である。彼が音声認識装置を作ろうとした時に，音韻モデルは作れただろうか？

5 神経生理学と音響的普遍構造

興奮性ニューロンの出力が抑制性ニューロンを介して近隣のニューロンを抑制したり，興奮性ニューロンへの入力が抑制性ニューロンを介して近隣のニューロンを抑制するなどの回路が生体において発見されており，反回性・順向性側抑制回路と言われる。ニューロン間の（ネットワークとしての）近さが刺激の空間的近さを表現していれば空間微分の演算子となり，時間的近さを表現していれば時間微分の演算子となり，空間的・時間的差異に反応する回路となる。例えば，網膜の視細胞は入力光の強度に反応するのではなく，その時間差分に反応する（固視微動と呼ばれる眼球の微動に対応して網膜への刺激を固定すれば世界は，10 秒で消失する）。このような差異（コントラスト）に着眼した処理は，時間的，空間的に隣接した刺激のみならず，離れた場所に存在する刺激間の差異に基づく処理も観測されている。また，差異に着眼する処理は，感覚器である網膜のみならず，中枢でも（視覚皮質の受容野特性など）広く観察される特性である。

このように視覚生理学においては，差異に着眼する処理は非常に多く観測されている。大脳皮質のコラム構造を発見したマウントキャッスルは視覚，聴覚などの領域を超えた普遍のアルゴリズムの存在を主張しており [10]，差異に基づく処理系が第一次聴覚野やそれ以降の処理に存在していても何ら不思議ではない。

6 まとめ

従来の音声科学・工学は［言語＋パラ言語＋非言語］から［パラ言語］（即ち，F_0 やパワー）を分離，正規化する方法論を検討してきた。親の声から［パラ言語］のみを分離し［言語＋非言語］を真似る幼児を，筆者らは見たことがない（自閉症児を除く）。［パラ言語］のみを分離する（不自然な）枠組みは，音声生成（調音音声学）に基づいて音声コミュニケーションを描いた結果得られる枠組みである。そろそろ異なる枠組みを検討すべき時期なのではないだろうか？

参考文献

[1] K. S. Scott *et al.*, Trends in Neurosciences, 26, 100–107 (2003)
[2] 柏野，言語，33, 9, 102–107 (2004)
[3] 峯松他，信学技法，SP2005-12，1–8 (2005)
[4] 峯松他，信学技法，SP2005-131, 121–126 (2005)
[5] 村上他，信学技法，SP2005-14, 13–18 (2005)
[6] 石井，文科省科学技術政策研究所，科学技術動向 45，13–24 (2004)
[7] R. D. Davis *et al.*, “The Gift of Dyslexia,” Perigee (1997)
[8] U. Frith, “Autism,” Blackwell Pub (1992)
[9] 東田直樹他，“この地球にすんでいる僕の仲間たちへ”，エスコアール (2005)
[10] V. Mountcastle, The Mindful Brain, MIT Press (1978)