

構造不変の定理とそれに基づく音声ゲシュタルトの導出

峯松 信明[†] 西村多寿子[†] 西成 活裕[†] 櫻庭 京子[‡]

[†] 東京大学 〒113-0031 東京都文京区本郷 7-3-1

[‡] 清瀬市障害者福祉センター 〒204-0013 東京都清瀬市上清戸 1-16-62

E-mail: [†]mine@gavo.t.u-tokyo.ac.jp, [†]nt-tazuko@ams.odn.ne.jp

[†]tknishi@mail.ecc.u-tokyo.ac.jp, [‡]sakuraba@mtd.biglobe.ne.jp

あらまし 『音声知覚の容易性と音声物理の多様性』 音声研究者を長年悩ませ続けてきた古典的問題である [1]。音声科学は、音素に対する音響的不変量の不在から、調音運動にその答えを求めている（運動理論）[2]。その一方で音声工学は、膨大なる音声データの収集とその統計的モデリングにその答えを求めている（隠れマルコフモデル）[3]。問題は解かれたのか？否である。本研究は、これらとは全く異なる方法で一つの解答を与える。その際、「構造不変の定理」と呼ぶ数学定理を導入する。この数学定理の上で、「何故言語（記号とその操作体系）は空気振動に宿ったのか？」「言語が宿るということは、空気振動に如何なる物理特性を要求するのか？」という問題を意識して音声モデリングを再考し、上記問題に対する解答、音声ゲシュタルト、を導出する。本稿では、この導出が、認知心理学、生態心理学、認知言語学、障害学、及び、複雑系研究を通して音声言語コミュニケーションを捉え直すことと等価であることを示す。更に、音声研究史に対する一つの提言を行なうと同時に、言語の獲得、及び、言語の起源についても考察する。

キーワード 音声知覚、音声認識、構造音韻論、構造不変の定理、音声ゲシュタルト、認知言語学、複雑系、自閉症

Theorem of the Invariant Structure and Its Derivation of Speech Gestalt

N. MINEMATSU[†], T. NISHIMURA[†], K. NISHINARI[†], and K. SAKURABA[‡]

[†] The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0031 Japan

[‡] Kiyose-shi Welfare Center for the Handicapped, 1-16-62, Kamikiyoto, Kiyose-shi, Tokyo, 204-0013 Japan

E-mail: [†]mine@gavo.t.u-tokyo.ac.jp, [†]nt-tazuko@ams.odn.ne.jp

[†]tknishi@mail.ecc.u-tokyo.ac.jp, [‡]sakuraba@mtd.biglobe.ne.jp

Abstract “Easiness of speech perception against great variability of speech acoustics.” This is the classical problem of speech science and engineering, which remains to be solved [1]. After the phonemes were found to be lacking in acoustically invariant features, speech science has pursued the solution in studies of the articulation (motor theory) [2]. Speech engineering, on the other hand, has pursued it in studies of statistical modeling of speech acoustics (hidden Markov models) using large speech corpora [3]. The problem is, however, yet to be solved. In this study, a completely different approach is taken to provide a candidate solution for the classical problem. The solution is obtained based on a mathematical theorem of the invariant structure. With the theorem, acoustic modeling of speech is reconsidered by asking another question “Why was language born in the aerial vibration?” or “What kind of physical properties of the aerial vibration was a necessary condition to the birth of language?” The reconsideration gives the candidate solution, called speech Gestalt. In this paper, it is also shown that the solution requires speech researchers to reconsider spoken language communication especially in view of cognitive psychology, physiological psychology, cognitive linguistics, disability, and complex systems. Further, a suggestion is made for speech science and engineering and some considerations are done regarding the acquisition of language and the origin of language.

Key words Speech perception and recognition, structural phonology, theorem of the invariant structure, speech Gestalt, cognitive linguistics, complex systems, autism

1. はじめに ～言語を言語たらしめるもの～

言語とは何か？身振り手振りなど様々な伝達手段がある中で、何故それは、音、即ち、空気振動の中に完成されたのか？空気振動の物理特性に、言語を創発する必要条件が揃っていたのか？

言語は脳が生み出したのか？とすれば、空気振動が異なる物理特性を持っていたとしても、言語は生まれたのであろうか？

「言語とは何か」という問いには様々な観点から答えられるが、本稿ではその問いを「言語学の第一課題とは何か」に置き換えて考えてみたい。人間は有限個の語を組み合わせることで、

無限の多様性を持つ文を生成できる。言語学の第一課題は「有限」から「無限」を生成する仕組の解明である[4]。無限の多様性と言っても、それは「カオスの複雑さ」ではなく「(凡そ)秩序ある多様性」を指す。これは如何にして生まれるのか？言語学の解答は「構造」である。文は句を単位とする構造、句は語を単位とする構造、語は形態素を単位とする構造、形態素は音節を単位とする構造、音節は音素を単位とする構造である、と言語学は説く[5]。文の上にも、段落、談話構造と更に大きな構造が存在し、また、言語を理解するためには意味構造に基づく知識も活用される。即ち、言語を階層的に捉えた場合、ある層における単位は、その下位層の単位(群)の構造として規定され、下位層の単位の実体として何を選択するのか(何が選択されるのか)の違いによって無限の多様性を生む。そして、その選択がある基準を満たすように行なわれるため、秩序ある無限の多様性が生成可能となる。但し、規則によって上位構造から下位構造へと導かれるのか(派生)、下位構造が下位構造同士の関係や制約に基づいて上位構造へと導かれるのか(併合・創発)、については言語学者の立場の違いにより大きく二分される。例えば初期の生成文法理論(生成音韻論を含む)は前者であり[6],[7]、近年の生成文法理論(ミニマリストプログラム)[6]、最適性理論[5]や、認知言語学(認知意味・文法・形態・音韻論)[7],[8]は後者である。文の構造を記述するとは、文が全体として成り立つような部分を指定し、その部分と全体の間に成立する構成的関係を必要かつ十分に指定することである[7]。

「簡単な規則群の適用を繰り返すことで派生される複雑さ」「下位層の単位(要素)間の相互作用・関係付けにより創発される、要素の総和以上の全体的な特徴」と言うのは複雑系研究における「複雑さ」の定義に相通じるものがある[9]。複雑系科学は、要素還元主義に基づく従来の科学のあり方に対するアンチテーゼとして興り、要素に還元できない全体としての系を分析・解明することを目的とする。相互作用によって関係付けられた要素群が、要素の総和以上の特性を有する全体を創発し、その全体がまた一つの要素となって更に大きな全体を創発する。その一方で、全体の特性が個々の要素の相互作用に影響を及ぼす。複雑系科学においては、このような階層構造として世界を捉える。クォークを要素として素粒子が、素粒子を要素として原子、そして分子が、分子を要素として生体高分子が、生体高分子を要素として細胞、そして個体が、個体を要素として社会が、といった階層構造を考える。下位層の要素群の関係付けに基づいて上位層の要素が構造として創発され、その繰り返しにより、階層構造が生まれる様子は上記した言語の階層構造と類似した考えであり、言語を複雑系としてモデル化する試みも、認知言語学において行なわれている[7],[8]。ここで、同質の操作を繰り返すことは複雑系では重要であり、自己相似性(フラクタル性)の創発などはその良い例である。本稿の目的は「このように複雑系との類似性も見られる言語が、何故、空気振動において完成されたのか」という問いに答える形で、言語としての音声物理の捉え方を提案することにある。言い換えれば、言語学を音響学、物理学の上で捉える試みである。それが、長きに渡る音声科学の古典的問題に対する一つの解答を与える。

2. 音声学と音韻論

音声学とは音声物理空間(音響空間或いは調音空間)を領域分割し、個々の単音と部分空間との対応を議論する。一方音韻論とは、個々の音を心的イメージとして捉え(単音から音素へ)、音素系列へと変換された発話の中に、どのような系列パターン・系列構造が存在するのかを議論する。物理イメージと心的イメージの差異はあるが、音声学は音声要素に分割し、音韻論はその要素群を構造へと統合する。問題は、音声物理と音素系列との対応が一对一にならず『音声知覚の容易性と音声物理の多様性』という古典的問題に落ちることである。

仮に音素が一对一対応となる物理特性を有していた場合、言語は空気振動に宿つただろうか？その場合、音素(音韻)構造として音節が創発されるメカニズムは自然発生したのだろうか？指を使えば数字が表現できる。指を「立てる・立てない」の2状態で考えれば、5本の指で $2^5 = 32$ 通りの、一对一対応がとれた音素の物理表現が可能である。日本語は十分表現可能である。なぜ、指の上に言語(即ち構造生成)は宿らなかったのか？

前節で下位要素群が構造を成し、やがてそれが階層構造を成すと説明した。構造化の際にもう一つ重要な点は「関係付ける要素群は時間的に離れた事象も含みうる」ことである。統語構造を考えれば、離れた語間の関係を捉えることが不可欠であることは自明である。所謂「長距離依存関係」である[10]。言語は局所的な関係付けだけでは生まれない。長距離関係を含む構造化・階層化が言語の本質であれば、言語を物理に落とす際に求めるべきは、音素と一对一に対応する物理事象ではなく、長距離関係を含む構造化を自然発生させる物理事象である。

音声コミュニケーションの目的を意味伝達と考えた場合、その目的達成に必要なのは語の情報(意味を持つ最小単位を形態素と考えれば、形態素の情報)を正確に伝えることである。もし、語を単位として音響的不変量が存在すれば、音素が不変量を持つ必要性は一切失われる。音素の定義を、意味論を切り離して言語学に求めた場合、2つの定義が導かれる[11]。1) a phoneme is a class of phonetically-similar sounds and 2) a phoneme is one element in the sound system of a language having a characteristic set of interrelations with each of the other elements in that system. 第一の定義が不特定話者音響モデルを導いたことは自明である。完全なる不変量は存在していなくても、凡そ、その不変量が存在し、後は統計モデルの構築で解決を試みた。しかし、話者適応技術の存在[12]は、それが未解決に終わったことを意味する。第二の定義は、音素は他の音素群と「ある関係」で結ばれた存在であることを主張する。更に第二の定義を推し進めると、The phonemes cannot be defined acoustically and they are a set of abstractions[11].となる。言語学は音素に音響的不変量など求めていないのである。本稿では、この第二の定義に基づいて「音と音の関係」を捉えることで、(凡そ)語を単位として音響的不変量が存在することを工学モデルに基づいて示す。更に、その不変量は長距離関係に基づく構造化を自然発生させること、即ち、空気振動において言語を自然発生させる可能性があることを数学的に示す。

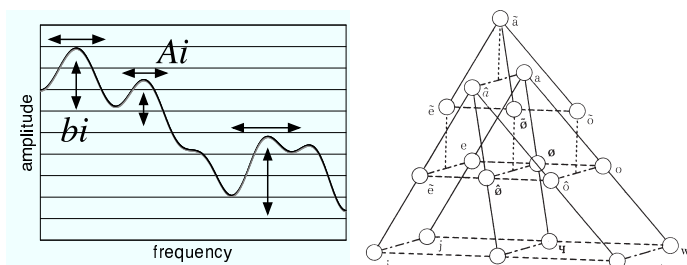


図1 二種類の不可避的な音響歪み 図2 構造音韻論と幾何学的音韻構造

3. 音声物理の多様性の工学的モデル化

何故、音声物理はかように多様なのか？それは不可避的に混入する非言語的特徴により音響特徴量が歪むからである。年齢・性別・体格といった発声者の個人性、収録系・伝送系・再生系といった音響機器特性、更には聴取者の聴覚特性にも個人差がある。Bark や Mel 尺度は平均特性に過ぎない。話すことは歪みであり、聞くことも歪みである。歪んでいない音声を入力する唯一の方法は「発声しないこと、聴取しないこと」である。

音声に混入する歪み・雑音は加算性雑音、乗算性歪み、線形変換性歪みの三種類に分類される。ここで加算性雑音は、雑音源の物理的の消去が可能であるという意味において不可避な雑音ではない（例えば、静かな部屋に移動すればよい）。一方、乗算性歪みや線形変換性歪みの消去は「発声しないこと、聴取しないこと」を要求するため、これらは不可避な歪みである。

乗算性歪みはマイクや伝送特性など、フィルタリングとしての歪みであり、話者性が GMM でモデル化されることを考慮すると、話者性の一部も乗算性歪みとなる。これらはケプストラムベクトル c に対するベクトル b の加算となる ($c' = c + b$)。

声道長差異（個体のサイズ差、年齢差）によるフォルマントシフトは、スペクトルの周波数ウォーピングとして捉えられる。また聴取者毎に異なる聴覚特性も、周波数ウォーピングとなる。連続かつ単調な周波数ウォーピングはケプストラム次元では行列 A を掛ける演算となる（線形変換性歪み）[13] ($c' = Ac$)。

即ち、不可避的な音響歪みは、スペクトルに対する縦方向 (b)、横方向 (A) に対する「ずれ」を生む作用素として捉えられる（図1）。複数の作用素 ($b_1, b_2, \dots, A_1, A_2, \dots$) による統合歪みも一次変換 $c' = Ac + b$ となり（即ちアフィン変換）、本稿ではこれを、不可避的な音響歪みの最も簡素な工学的モデルとして採用する。なお、本モデルは大局的な MLLR や SAT に相当する。triphone の場合、物理的に異なる状態は異なる話者群の音声によって学習されるため、状態毎にその話者性は異なる。そのため、単一話者への適応は複数の変換行列を要求することになる。とは言え、大局的な変換行列だけであれば、非言語的特徴の表現能力は低減する。この対処については後述する。

4. 構造不変の定理と構造音韻論の物理実装

4.1 構造音韻論と構造主義

ソシュールの言語哲学「言語は概念的差異と音的差異だけである」[14] に啓蒙され、ヤコブソンによって構造音韻論と呼ばれる言語学の一分野が確立された。図2にフランス語母音及び

準母音に対する幾何学構造を示す[15]。母音間の線分形状の違いは対応する二母音間の弁別素性的な差異を表現する^(注1)。また、ハレはロシア語音素に対して弁別素性を用いた音素樹型図を構成している[16]。ここでは、各部分木が自然類となるように、素性を各ノードに割り当て、言語学的に妥当な音素分類木を構築している。両者の音素群表象の裏に「構造」の意識があることは自明であり、この場合、音素は他の音素との関係において初めて意味を持つ存在となる。即ち、第2.節における第二の音素定義である。構造音韻論では、母語話者の音声には性別、年齢、話者を問わず、同一の音韻構造が普遍的に存在すると主張する。即ち、音声に不可避な非言語的特徴は、言語学者の頭の中で消失される。本稿では彼らの心的認知過程の数学的解釈を行ない、音韻群の普遍的な幾何学構造を物理実装する。

なお、ソシュール、ヤコブソンと引き継がれた構造の意識は、レビストロース（人類学）に受け継がれ、構造とは「ある数学的変換に対して不変な特性」として定義されることになる[17]。

4.2 構造不変の定理による構造音韻論の物理実装

ケプストラム空間において音素を点で表現すれば、 n 音素群は n 点から成る幾何学構造として表現される。第3.節に示したように、不可避的な非言語的特徴はアフィン変換となるが、この変換は構造を歪ませる変換として広く使われている。即ち、 n 点構造は非言語的特徴により不可避的に歪むことになる。[13]において導出された変換行列には、回転（鏡像も含む）や遷移以外の変形要素が含まれている。結局音韻構造は、数学的には、必ず話者・収録環境・聴取者間で歪まざるを得ない。本稿ではこれを、歪みの無い構造として捉える枠組みを検討する。

不可避的に歪む構造を、構造不変な枠組みとして捉えるにはどうしたらよいのか？答えは単純・明快である。構造が歪まないように、空間を歪めて表象すればよいだけである。空間を歪ませることを許可した場合でも、与えられた任意の歪んだ構造に対して、どのような空間の歪みを導入すればよいのか？これらの問いを下記に示す「構造不変の定理」と呼ぶ数学的事実を導入することで解決する。なお「空間を歪ませて観測する」という方法論は、重力の発生を説明するために時空を歪ませて議論した一般相対性理論と、数学的には等質である。

空間内に存在する n 点に対して、 ${}_nC_2$ 個の全ての二点間距離を求めると、その n 点で張られる構造は一意に規定される。

構造不変の定理：意味のある記述が分布（ガウス混合分布）としてのみ可能な物理現象を考える。分布群に対して、全ての二分布間距離を求める（距離行列）。二分布間距離として、バタチャリヤ距離、カルバック・ライブラ距離、ヘリンガー距離などを用いた場合、各分布に対して単一の任意一次変換を施しても、二分布間距離は不変である。即ち距離行列は不変であり、その結果、構造も不変となる（図3参照）。

バタチャリヤ距離を用いて話を進める。本距離尺度は、二つの

(注1)：弁別素性は本来、ソシュールの言う音的差異を表現するために考案された。やがて全ての音素間差異群を満たすように、ヤコブソンは「素性の束」として音素を絶対的に定義した。これは第2.節における音素の定義を第二の定義から第一の定義へと移すことに相当し、その意味において、音素を「素性の束」として捉える音韻論をヤコブソンの「勇み足」として捉える言語学者もいる。

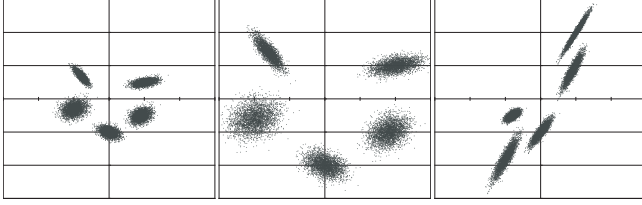


図3 構造不変の定理（これらが全て同一構造となる）

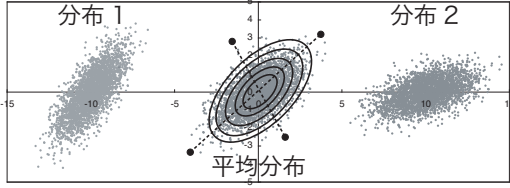


図4 平均分布形状とそれに基づく空間の歪み

確率密度分布に対する正規化相互相関の $-\ln$ と等価であり [18], 二分布間の「関係」を「距離」として定義している^(注2)。なお, BD 距離の単位であるが, $0 \leq \int_{-\infty}^{\infty} \sqrt{pq} dx \leq 1.0$ を確率として解釈すれば, 下式は自己情報量, 即ち単位は [bit] となる。

$$BD(p(x), q(x)) = -\ln \int_{-\infty}^{\infty} \sqrt{p(x)q(x)} dx \quad (1)$$

上記定理に示すように, 二つの分布が混合ガウス分布である場合, 任意の単一アフィン変換によって分布間距離は不変である。以下, 本定理と一般相対性理論との数学的接点について述べる。単一ガウス分布間のバタチャリヤ距離は下式となる。

$$BD = \frac{1}{8} \mu_{pq}^T \left(\frac{\sum_p + \sum_q}{2} \right)^{-1} \mu_{pq} + \frac{1}{2} \ln \frac{|\sum_p + \sum_q|/2}{|\sum_p|^{1/2} |\sum_q|^{1/2}} \quad (2)$$

第一項は平均ベクトル差異に基づく距離, 第二項は分布形状差異に基づく距離である（両者ともアフィン変換不変）。第一項は, 両分布の分散共分散行列の平均行列によって計算される平均分布形状に対するマハラノビス距離である。即ち図4に示すように第一項は, 二つの平均ベクトルの位置関係（方向）に依存させて, 単位距離を変化させている。これは分布間距離を, 両分布の位置・形状に基づいて空間を歪ませて求めることに等しい。空間を歪ませて距離尺度を定義する場合, 非ユークリッド幾何学であるリーマン幾何学（微分幾何学の一様）を用いて解析することが可能である。説明を簡単にするために, 一次元の単一ガウス分布 $p(x)=p(\mu, \sigma)$ を考える。この時 BD は,

$$BD(\mu_p, \sigma_p, \mu_q, \sigma_q) = \frac{1}{4} \frac{(\mu_p - \mu_q)^2}{\sigma_p^2 + \sigma_q^2} + \frac{1}{2} \ln \frac{\sigma_p^2 + \sigma_q^2}{\sigma_p \sigma_q} \quad (3)$$

となる。平均 μ , 標準偏差 σ の二次元空間を考え（ガウス分布はこの空間上の点となる）, $d(BD)$ をその計量を用いて

$$d(BD) = M_{\mu\mu} d\mu^2 + 2M_{\mu\sigma} d\mu d\sigma + M_{\sigma\sigma} d\sigma^2 \quad (4)$$

(注2): N 個の値をとる二つの確率変数 p_i, q_i を考える。 N 次元ユークリッド空間において点 $(\sqrt{p_1}, \dots, \sqrt{p_N})$ と点 $(\sqrt{q_1}, \dots, \sqrt{q_N})$ が成す角を θ とすると, $\cos \theta = \sum_i \sqrt{p_i} \sqrt{q_i}$ であり, これは正規化相互相関と等価である。これを対数化した $-\ln(\cos \theta)$ が BD であり, 確率分布を離散分布から連続分布（密度関数）にしたものが式(1)である。この場合, N 次元空間は ∞ 次元空間へと拡張され, 全ての確率密度関数はこの超空間の単位球上に存在する点として解釈される。

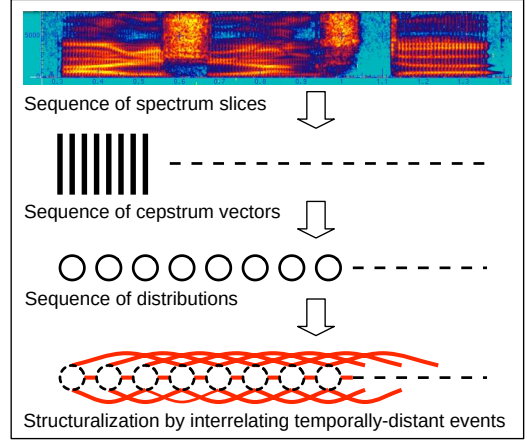


図5 単一発声の構造化

とすると, 計量 $M_{\mu\sigma}$ は以下のように算出される。

$$M_{\mu\sigma} = \frac{1}{8} \int_{-\infty}^{\infty} p \frac{\partial \ln p}{\partial \mu} \frac{\partial \ln p}{\partial \sigma} dx \quad (5)$$

定数項を無視すると, これはフィッシャー計量（情報計量）と呼ばれる量である。パラメトリックな確率密度分布をそのパラメータが張る空間の一点として捉え, フィッシャー計量を用いた場合の空間は, 情報幾何学 [19] などにおける多様体である。即ち第3.節で示した工学モデルに基づく構造音韻論の物理実装は, 多様体と呼ばれる歪んだ空間において実現され, これを音響的普遍構造と呼ぶ [20]。この時, A を掛ける演算は構造の回転として, b を足す演算は構造のシフトとして解釈される。

なお「構造不変の定理」は 1) 事象を記述するための分布関数, 2) 事象に対する変換関数, 3) 二事象（分布）間の距離関数の組み合わせにおいて成立する。本稿で示した, 1) ガウス混合分布, 2) アフィン変換, 3) バタチャリヤ距離以外の組み合わせにおいても成立する可能性がある。特に変換関数に対して非線形な関数を導入した場合の本定理の成立条件などは興味深い。

5. 単一発声の構造化と音声ゲシュタルト

5.1 「構造不変の定理」の単一発声への適用

構造音韻論の物理実装という形で音響的に普遍かつ不変なる構造の存在を数学的に導いた。しかし本定理は, 何らかの言語単位を要求するものではない。対象とする事象が分布として記述されることのみを要求する。その意味において, 本定理は単一の発声に対しても適用可能である。その様子を図5に示す。音声物理は常に揺れている。同一のピッチ波形を繰り返したものはブザー的な音になる。この事実を考慮すれば, スペクトル系列を分布系列として表象することは音声物理の特性に適した表現と言える。対象とする音声区間に存在する全分布に対し, 全ての二分布間距離（即ち関係）を求めると, それは一つの構造を規定する。そしてこの構造は, 非言語的特徴に対して不変である。 A を個体の成長に伴う音響歪み, b を個体の移動に伴う音響歪みと解釈すれば, 本定理は, 長距離関係を含む, 事象間の全関係を通して対象の全体的構造を捉えることで, あらゆる個体があらゆる場所において「普遍・不変なる物理事象」即ち「シンボル（記号, 語）」を創発できることを保証する。

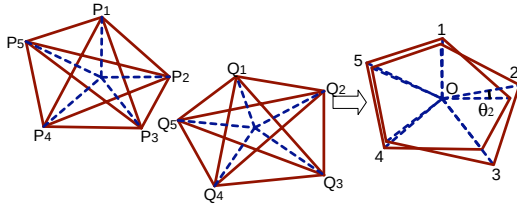


図 6 回転とシフトによる構造の移動と照合

5.2 ユークリッド空間における構造間距離尺度

N 次元ユークリッド空間において、構造（距離行列）として与えられた二発声間の距離尺度を考える。なお、対象とする二発声が同数（ M ）の分布系列として表象されているとする。この時距離尺度は、一方の構造をシフト及び回転させて他方の構造へと近づけ、対応する二事象間距離の和の最小値を求める問題となる。ユークリッド空間において回転だけを許す行列は直交行列であり、二事象間距離の和の最小値は下記で算出される。

$$\sum_{i=1}^M (\overline{OP_i}^2 + \overline{OQ_i}^2) - 2 \sum_{i=1}^M \sqrt{\alpha_i} \quad (6)$$

O は両構造の重心である（構造を移動し重心を重ねる）。 α_i は N 次正定行列 $S^t T T^t S$ の固有値である。 S は行列 $(\overline{OP_1}, \dots, \overline{OP_M})$ であり、 T は行列 $(\overline{OQ_1}, \dots, \overline{OQ_M})$ である。なお、この距離尺度は「構造不変の定理」が規定する構造には直接的には適用できない。「構造不変の定理」は非ユークリッド空間を要求しており、三角不等式が必ずしも満たされないからである。なお、上記構造間差異の近似として $\sum (\overline{OP_i} - \overline{OQ_i})^2$ を考えた場合、これは距離行列をベクトルとして見なした場合のユークリッド距離で近似できることは [21] に示した通りである。

以上、 A （回転）及び b （シフト）に基づく音響モデルの適応処理を施した後の音響スコアが、 A や b を直接的に算出しなくても、かつ、音声の物理的実体の情報を直接的に用いなくても計算可能であることを示した。これは、各音の同定（絶対音感）が出来なくても音楽が楽しめるように、各音の同定をすることなく音楽が楽しめることを示唆している。語の同定さえ出来れば、心的辞書の該当項目を参照することで、その語を構成する音素列の情報は、事後的かつ自動的に既知情報となる。

5.3 音声の構造的表象に対する認知・生態心理学的解釈

音楽を聞く時に、一つ一つの音を同定できなくても、音楽を楽しみ、その音楽（の名前）を同定することができる。これは、音楽に不可欠な能力が音程の変化を相対的に捉える能力であり、絶対音感とは音楽に本質的に必要とされる能力ではないからである [22]。ある曲を移調した場合、個々の音の物理特性は当然変化するが、音程遷移は不変なため、移調前後の曲において同一メロディーを知覚できる。エーレンフェルスは音楽のこのような特性（個々の要素の総和以上の全体的特徴であり、また、ある変換操作に対して不変性を有する事象）をゲシュタルト質と呼んだ [23]。ゲシュタルト知覚は様々な知覚現象として観察されている。有名な例は錯視である。図 7 左にティチエナー錯視と呼ばれる錯視の例を挙げる。当然のことながら、二つの中心円以外のパターンが存在しなければ、これらは全く同一サイズに見える。即ち錯視は、一つ一つの要素の知覚では説明不可能な、部分の総和以上の全体的な知覚によって誘発されると解釈

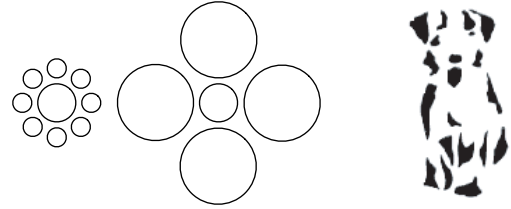


図 7 視覚におけるゲシュタルト知覚の例

できる。特に図 7 左に示したような幾何学的錯視は「図形群の形状、配置、関係に応じ、その周囲の空間を歪ませて知覚する現象である」と解釈できる。実際に「重力モデル」と呼ばれる幾何学的錯視の説明モデルでは、アインシュタイン方程式を直接的に解き、シュバルツシルト解を求めて歪んだ空間を規定する形で、種々の幾何学的錯視を説明している [24]。

提案する音声の物理表象は、個々の音の物理的実体を全て捨て去り、要素間の相互作用に基づき音響空間を歪ませる形で、その全体的構造のみを表象している。かつその構造は非言語的特徴（アフィン変換）に一切不変である。以上の特性を考慮し、提案する音声の構造的表象を音声ゲシュタルトとして定義する。

ゲシュタルトはエーレンフェルスによって提唱されたが、同様の知見はマッハ（物理学・心理学）によって既に提唱されており、エーレンフェルスはマッハの影響を受けていると言われている [23]。マッハは物理学研究において「ある物体の質量は、その物体の周りの全ての物体との関係で決る。他に何も無い空間の中では、ある物体の質量には何の意味もない（マッハの原理）」と述べ [23]、物理的世界における事象と事象との相互作用・関係の重要性を説いた。一方で心理学研究においても「全ての感覚は、感性的諸要素の関数的複合体である」と述べ、心理的世界における事象と事象との相互作用・関係の重要性を説いている [25], [26]。また「空間・時間が人間の感覚である以上、その生態学的空間・時間は非ユークリッド的である」と述べ [23], [25]、更に物理的空間・時間における非ユークリッド性をも指摘している [26]。周知のようにアインシュタインはマッハの影響を多分に受けており、アインシュタインによれば「マッハは空間をあらゆる質点間の瞬間的な相対距離の総和という考えでもって置き換えようとした」[23] のであり、宇宙を一つの構造として解釈していることが分かる。複雑系の議論を考慮するならば、マッハは「物理的世界の複雑性」と「心理的世界の複雑性」の両者を指摘しており、その先見の目は驚嘆に値する。

さて、第 4.1 節においてソシュール（即ち言語学）に端を成す構造主義について述べた。マッハに端を成す物理学・心理学における構造の意識も一つの構造主義として捉えることができ、ゲシュタルト心理学は心理学の構造主義とも言われる。ソシュールとマッハは全く同時代の先駆者であり、本稿で提案する手法は「両者の主張を融合する形で導出される音声の物理表象」との位置づけを持つ。近年、ゲシュタルトなど、人間の基本的な認知特性に基づいて言語活動を捉え直す言語学として、認知言語学が注目を浴びているが、その一分野である認知音韻論は「基底形のような抽象的な存在でも規則や制約でもなく、ゲシュタルト単位としての単語全体、即ち実際に発話され、まとまった単位として全体的に知覚される単語」[7] を研究対象とす

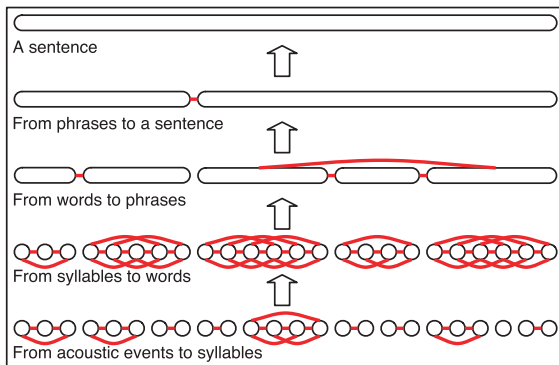


図 8 物理から言語へ

る。しかし現在の認知音韻論研究は、物理的な音ではなく、心理的な音（音素系列）が研究の出発点となるため、音から音素（物理から言語）への認知過程については何も言及しない^(注3)。本研究は、言語の本質とも言える「構造」を前提として音声物理を捉え直す試みであり、認知音韻論を「言語が生まれた場」即ち「空気振動」へと導くことも筆者らの狙いの一つである。

ゲシュタルト心理学に端を発し、ギブソンはアフォーダンスという概念を提唱した。「環境が提供する行為の可能性」と解釈される概念であるが、その概念獲得に至る視覚研究において、本研究の構造的表象と類似した議論を見ることができる。ギブソンは視覚の本質が「光の差異が成す構造」「動きの中に見られる不変項」にあることを指摘している[27]。提案している音声の物理表象は「ランニングスペクトルの中から、音的差異のみを抽出し、全差異によって構成される普遍かつ不変な物理構造」と定義されることから、本構造的表象はアフォーダンスとしての側面も持つと言える。そして、その構造によって「シンボル」が創発され、やがてそれが何らかの対象（意味）と恣意的に結びつくと考えれば、空気振動という環境の中に音声アフォーダンスを見いだすことで、人間は言語（無限なる構造的多様性）を獲得した、との仮説を立てることも不可能ではない。

認知言語学では、語をエージェントとして仮定し、文をマルチエージェント・システム（語群の相互作用が成す複雑系）として捉える試みもある[28]。このように考えると、音構造が音節を、音節構造が語を、語構造が句を、句構造が文を、と物理から言語に至るまでを「要素間の相互作用・関係」に基づいて再帰的に構築することができる（図8参照）。[10]には「音韻構造と意味との連合が最初にあり、統語構造は、意味構造と音韻構造を調節する補助として後になって進化したと考えても完全に理にかなう」と述べられている。生成文法論者が重要視する文構造の再帰性[29]も、複雑系の観点から考えれば、自己相似性（フラクタル性）の一形態として捉えることができる。

6. 幾つかの実験的事実とそれに基づく提言

6.1 話者性可変音声の聴取[30]

互いに移調関係にある二つの曲を用意する。この曲の間で音符の交換を行えば当然異なるメロディーとなる。話し手と聞

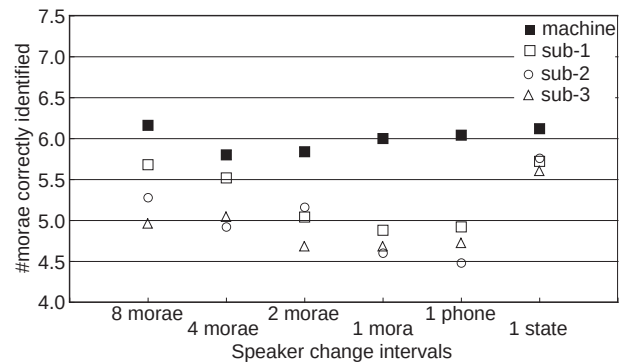


図 9 非音声研究者によるモーラ同定率

き手との間に張られるコミュニケーションチャネルは、音声ゲシュタルトにより、音響歪み無しの情報伝達が可能となる。しかしこれは、非言語的特徴を表現する *A* と *b* が時不変であるとの仮定がある。HMM 合成を用いれば、上記メロディーの例の様に、話者性を時変にできる（話者性可変音声）。音声ゲシュタルトは長距離関係に基づく構造として定義されるため、話者性可変音声は、その構造を必ず歪ませることになる。このような音声を人間は正しく聴取できるだろうか？話者性変化の一部を音韻の変化として捉えてしまうことはないだろうか？

「無意味 8 モーラ列の書き取り」をタスクとして選んだ。男性アナウンサー 7 名に対する合成用の HMM を話者毎に学習した。話者性変化のタイミング制御として、8 モーラ（話者性固定）、4、2、1 モーラ、1 音素、1 分布の 6 段階を用いた。分布単位での話者性変化は、1 音素当たり 5 回話者が変わる。合成音の F_0 パターンは、LHHHLLLL（第 4 型）に固定した。話者性変化の各タイミング（全 6 種類）に対して、ランダムに選ばれた 8 モーラ列を 25 種類用意し、最終的に 150 個の無意味 8 モーラ合成音声を作成した。なお合成音声の品質、及びモーラ同定の難度を考慮し、促音、撥音、拗音、濁音、半濁音を除いて刺激音声を作成した。用いたモーラ種類数は 43 である。

被験者は成人男性 8 名であるが、内 5 名は音声研究に従事する者であり、他 3 名は今回初めて合成音声の聴取実験に臨む被験者である。刺激呈示は 2 回ずつ、ヘッドフォンを通して行なわれ、PC 上で書き取る形をとった。呈示音声が 8 モーラであること、一部のモーラが含まれないことは事前に教示した。

聴取実験の後、全ての合成音声を、性別非依存不特定話者音響モデルを用いて連続モーラ認識させ（但し系列長は 8）、音声認識器（HVite v3.2.1）のモーラ同定率も算出した。

音声研究者の場合、1 人を除き、話者性変化による同定率の有意な低下 ($p=0.1$) は観測されなかった。一つ一つの音を絶対的に同定する方策（即ち分析的聴取）が採られていると解釈される。図 9 に音声認識と非音声研究者の結果を示す。音声認識も話者性変化に因らず安定したモーラ同定率を示している。性別非依存不特定話者モデルを用いているため、話者性の違いは統計モデルに吸収されていると解釈できる。一方非音声研究者の場合、話者性変化による同定率の有意な低下が観測された。sub-1 における 8m-2m ($p=7.54\%$)、8m-1m ($p=3.56\%$)、8m-1p ($p=5.46\%$)。sub-2 における 8m-1m ($p=6.04\%$)、8m-1p ($p=1.58\%$)、2m-1p ($p=5.81\%$)。なお m はモーラを表し、

(注3)：認知音韻論と言えども言語学である以上、物理現象への言及は不十分であるように思われる。逆に音声工学は言語への言及が不十分であると考えられる。

表 1 LPF の cut-off 周波数と認識精度

cut-off [kHz]	8.0	4.0	3.5	3.0	2.5	2.0
accuracy [%]	43.0	62.8	81.8	96.9	80.0	100.0

p は音素を表す。非音声研究者の場合、個々の音の情報よりも、音声全体の特徴に基づいて処理が行なわれていると示唆される（非分析的聴取）。なお、分布単位で話者を変化させた場合に同定率が向上しているのは、HMM 合成はスペクトル生成時に平滑化処理を行なうため、話者性変化間隔が極端に短いと、話者性変化が表出されなくなるためであろう。次節で示す認識実験結果を踏まえ、本実験結果について第 6.3 節で再度考察する。

6.2 日本語母音系列の自動認識 [31]

第 5.2 節において、音声の構造的表象を用いた音響照合を論じた。提案表象は全ての音声事象を相対的に捉えるため、構造の回転により異なる 2 語を等しいと判定する可能性もある。つまり構造が単純過ぎると認識能力は劣化する。以下、日本語 5 母音の孤立発声系列を用いて認識性能を検討した。対象語彙は V_1 - V_2 - V_3 - V_4 - V_5 ($V_i \neq V_j$) であり、語彙数は ${}_5P_5=120$ である。

母音の孤立発声系列（系列長 5）より、各母音の中心部のみから分布を推定し、距離行列を算出する。少量データを用いた分布推定となるため、ここでは事後確率最大化基準に基づいて分布パラメータの推定を行なった。5×5 の距離行列の上三角部分をベクトルと見なして、これを構造ベクトルと定義する。一つの母音系列は一つの構造ベクトルへと変換される。この構造ベクトルを用いて 120 の語彙の統計モデルを多次元ガウス分布として構築し、それを単語モデルとして音声認識を行なった。

学習用話者は成人男性 1 名であり、評価用話者は学習用話者とは異なる成人男性 4 名、女性 4 名である。音声ゲシュタルトは非言語的特徴を表現する次元を失った物理表象であるため、学習用話者は 1 人である。第 3. 節に示した様に、一次変換は最も簡素な非言語的特徴のモデルであるため、その記述能力に限界があることは容易に推測される。そこで非言語的特徴をより効果的に消失させるための前処理を検討した。[32] は「話者性の多くはパワースペクトルの 2.2kHz 以上の帯域に観測される」との分析結果を呈している。これに倣い本実験では、前処理として低域通過フィルタ（LPF）を施すことを検討した^(注4)。

実験結果を表 1 に示す。語彙数が 120、即ちチャンスレベルが 0.8%であることを考慮すると、cut-off 周波数が 8.0kHz 時においても相対的には十分な認識能力があると言えるが、絶対的性能は低い。しかし、cut-off の低下に伴い認識率は向上し、2.0kHz 時に 100%に達した。即ち、個々の音の物理的実体（フォルマント周波数やスペクトル形状）を直接的に用いなくても、音声の全体的な特徴（音声ゲシュタルト）のみで、5 母音系列は完全に認識可能であること、しかも、学習話者は 1 人で十分であることが示された。cut-off と認識率が単調な関係に無いのは、cut-off を上げるとより高次のフォルマントまでを含む帯域が利用できるようになる一方で、話者性もより強く表出されるため、そのトレードオフに因るものと考えられる。

6.3 音声ゲシュタルトはどこで生まれるのか？

2 つの実験結果を音楽の聴取と比較して考察する。聴取者を絶対音感保有者とそれ以外に区分して考える。音楽は、相対的な音高関係の意識確立がその本質であり、相対的変化（パターン）の知覚を通して音楽を楽しむ [22]。しかし筆者らの予備調査によれば、絶対音感者は音程変化パターンを知覚している自覚がない。彼らは全ての音が音符（シンボル）に変換されるため、パターン知覚を自覚できない。更には「音をシンボル化出来ない人の音楽聴取が想像できない」と意見する。この議論を音声に対して行なう。母語であれば、聴取者は全ての文発声を音素（シンボル）列に変換できる。ここで、絶対音感者と同様に「音声を相対的に聴取しているにも拘らず（音声ゲシュタルト）、それを自覚できないだけ」と考察出来ないだろうか？

相対音感すら保有しない聴取者を無音感者とした場合、音声聴取における無音感者は存在するのだろうか？音韻の生成、及び同定が出来ない人間である。筆者らが考えるに、それは胎（乳）児である。彼らは喉頭が下がっていないため、共鳴スペースを持たず、物理的に調音できず、共鳴音が作れない。口腔（共鳴スペース）は、二足歩行による喉頭の下落に起因するため人間特有の形質であり、「人間は猿として生まれる」と言える [33]。更に音韻レパートリーを知る由も無く、音韻同定は不可能である。人間と言語との出会いは生前約 3ヶ月前から始まる「調音無き聴取」^(注5)であることは普遍的真実である [34], [35]。音を出せない、音を同定できない胎（乳）児が音声をどのように聞くのだろうか？調音指令に変換して聞いているのだろうか？「あいうえお」を分節して母音同定しているのだろうか？図 7 右に示す図を、個々の要素に分割した後で全体を認知する人は、通常いない。個々の要素は無意味だからである。胎（乳）児にとって「あ」は無意味である。音楽の聴取と同様に、音声の聴取もその全体的な特徴をゲシュタルトとして知覚すると考えるのが最も自然な解釈であろう。音声の全体的特徴を考える時、従来、韻律ばかりを捉えてきた。超分節的特徴＝韻律的特徴という図式である。音声ゲシュタルトは分節的特徴と言われるスペクトルのみを使って超分節的特徴を定義している。そして、その音声ゲシュタルトのみを使って日本語 5 母音系列は完全に認識できる。前節では 2kHz の LPF により認識率が 100%になったが、[35] は「唇から子宮内の羊水^(注6)」に至る経路が凡そ、2kHz の LPF となることを示している。即ち前節の実験は、胎児が子宮内で聞く母親の声を標準パターンとすることのみで、不特定話者による日本語 5 母音系列は認識可能であることを示唆している。しかも個々の母音の知識を一切持たずに、である。

従来の音声研究パラダイムでは、音声ストリームを要素に切り刻む方法論に終始していた。これはどの程度正しいのだろうか？人間は「聴取」を通して言語と遭遇する。しかし、科学は「発声」を通して言語と遭遇した。[36] も音声の「受容」を描く音声学の不在を指摘している。即ち従来のパラダイムでは人間と言語との遭遇は描けないことになる。調音音声学、運動

(注4)：本来の目的は、高域において音素間のスペクトル差異を消失させることが目的であり、高域のスペクトル成分を落とすことが目的ではない。

(注5)：この頃までに、形態学的に聴覚は完成する。

(注6)：実験的には、水で満たした胃と口前にマイクを設置している。

理論, ソース・フィルタモデル, 分節的特徴と韻律的特徴, 全てが「発声」を通して得られる帰結である。そもそも, speech という単語が what is spoken として定義される以上, speech science では, 人間と言語との遭遇は描けないのかもしれない。

6.4 障害学研究から見た音声ゲシュタルト [37]

音声ゲシュタルトに基づく知覚が音声コミュニケーションの根幹に存在していると仮定した場合, 音声をゲシュタルトとして認知できなければ, どのような状況に置かれるのだろうか?

ある障害を持つ方々について次のことが知られている。図 7 のような錯視が起こり難い。刺激群を「まとまり」として捉えることが苦手である。絶対音感を持つ可能性が高い。マガーク効果が起こり難い。物事の間隔を捉えることが苦手である。逆に, 関連性の薄い情報を覚える能力に長けている。時刻表, 電話帳, 漢和辞典を暗記した例が報告されている。物事の抽象化・概念化が苦手である。その障害とは自閉症である。自閉症者は音声言語によるコミュニケーションが苦手であり, 一生言語を獲得できない者もいる。視覚刺激の方が理解し易く, 8~9 割は視覚に頼った生活をしている, と言われる。筆者らは自閉症者の音声知覚と音声ゲシュタルトとの接点について実験的な結果を有してはいない。関係者との議論を開始したところである。

音声の物理空間を分割し, 独立性を仮定して各要素をモデル化する。音響モデリングである。自閉症者は 60 億の [a] を覚えようとしているのかもしれない。でも次の瞬間に, 新たな [a] と遭遇する。かつてロボット研究者は, フレーム問題を通して「日常世界の複雑さ, 時々刻々起こる新規性」に気付かされた。その複雑な世界を処理する人間の認知特性も, マッハが指摘したように, 複雑である。未知なる環境に立ち往生するロボットと, 環境の変化に立ち往生する自閉症者の類似性が報告されている [38]。環境からの情報 (アフォーダンス) を適切に取得できていないのである。「音声物理の複雑さ・多様さ」への対処は, 処理系を複雑系にすることで達成されると筆者らは考える。

7. まとめ ~空気振動を言語たらしめたもの~

音声物理の要素分割に基づく従来の音響的要素モデリングに対し, 複雑系としての言語, 及び, 複雑系としての認知特性を重ね合わせ, 音声物理事象の相互関係だけを捉える関係モデリングを提案した。また, 幾つかの実験事実を通してその妥当性を論じると共に, 言語の獲得, 起源についても考察した。今後, 相互関係モデリングをより精緻化すると共に, 相補的に位置付けられる二つのパラダイムの融合について検討する予定である。

本稿をまとめるに当たり, 一つの思考実験をしたい。かつて音声科学者が求め続けた「音素と一対一対応する物理特性」が存在した場合, 果たして言語は空気振動に宿っただろうか? 筆者らの答えは「否」である。何故なら, 長距離関係に基づく構造化が自然発生する可能性を見失うからである。人間同様, タマリン猿も 2 つの言語音の識別ができるが, 彼らには言語が無い。[39] は, 言語音の「差異」を検知する能力だけでは不十分であり, 言語音「差異」を coordinate する能力が必要であると説いている。当然その物理的解釈についての言及は無いが, 本稿はその説に対する一つの数学的解答を与えている。

文 献

- [1] K. Johnson and J. W. Mullenix, *Talker Variability in Speech Processing*, Academic Press (1997)
- [2] 柏野牧夫, “音声知覚の運動理論をめぐって”, 音講論, 1-2-10, pp.243-246 (2004)
- [3] 徳田恵一, “隠れマルコフモデルによる音声認識と音声合成”, 情報処理, vol.45, no.10, pp.1005-1011 (2004)
- [4] 窪蘭晴夫, “音韻論”, 岩波講座・言語の科学・音声, 第 2 章, 岩波書店 (1998)
- [5] 窪蘭晴夫, 音声学・音韻論, くろしお出版 (1998)
- [6] 北川善久他, 生成文法の考え方, 研究社 (2004)
- [7] 吉村公宏編, 認知音韻・形態論, 大修館書店 (2003)
- [8] 辻幸夫, ことばの認知科学事典, 大修館書店 (2001)
- [9] 金子邦彦他, 複雑系のカオスのシナリオ, 朝倉書店 (1996)
- [10] レイ・ジャッケンドフ, 心のパターン, 岩波書店 (2004)
- [11] H. A. Gleason, *An introduction to descriptive linguistics*, New York: Holt, Rinehart & Winston (1961)
- [12] 篠田浩一, “確率モデルによる音声認識のための話者適応技術”, 電子情報通信学会論文誌, D-II, vol. J87, no. 2, pp.317-386 (2004)
- [13] M. Pitz *et al.*, “Vocal tract normalization as linear transformation of MFCC,” *Proc. Eurospeech*, pp.1445-1448 (2003)
- [14] フェルディナン・ド・ソシュール, 一般言語学講義, 岩波書店 (1940)
- [15] ローマン・ヤコブソン, 構造的音韻論, 岩波書店 (1996)
- [16] M. Halle, *The sound patterns of Russian*, The Hague: Mouton (1959)
- [17] 橋爪大三郎, はじめての構造主義, 講談社現代新書 (1988)
- [18] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of Calcutta Maths Society*, vol.35, pp.99-110 (1943)
- [19] 甘利俊一他, 情報幾何学の方法, 岩波書店 (1993)
- [20] 峯松信明他, “構造不変の定理に基づく音声の構造的表象とその距離尺度”, 音講論, 1-5-13, pp.25-26 (2005)
- [21] N. Minematsu, “Yet another acoustic representation of speech sounds,” *Proc. ICASSP*, pp.585-588 (2004)
- [22] 宮崎謙一, “「絶対音感」はどこまで分かったか?», 日本音響学会誌, vol.60, no.11, pp.682-688 (2004)
- [23] エルンスト・マッハ, 時間と空間, 法政大学出版局 (1977)
- [24] 後藤卓男, 田中平八, 錯視の科学ハンドブック, 東京大学出版会 (2005)
- [25] エルンスト・マッハ, 感覚の分析, 法政大学出版局 (1971)
- [26] エルンスト・マッハ, 認識の分析, 法政大学出版局 (2002)
- [27] 佐々木正人, アフォーダンス-新しい認知の理論, 岩波書店 (1994)
- [28] K. Kuroda, *Foundations of pattern matching analysis - a new method for the cognitively realistic description of natural language syntax*, Dr. thesis of Kyoto University (1996)
- [29] M. D. Hauser, N. Chomsky, W. T. Fitch, “The faculty of language: what is it, who has it, and how did it evolve?” *Science*, vol. 298, pp.1569-1579 (2002)
- [30] 峯松信明他, “構造音韻論の物理実装に基づく新しい音声の音響的表象”, 信学技法, SP2004-27, pp.47-52 (2004)
- [31] 村上隆夫他, “音声の構造的表象を用いた日本語母音系列の自動認識”, 信学技法, SP2005 (2005-5, 発表予定)
- [32] T. Kitamura *et al.*, “Speaker individualities in speech spectral envelopes,” *JASJ(E)*, vol.16, no.5, pp.283-289 (1995)
- [33] 葉山杉夫, ヒトの誕生, PHP 新書 (1999)
- [34] 正高信男, 子どもはことばをからだで覚える, 中公新書 (2001)
- [35] I. Yamanouchi *et al.*, “The transmission of ambient noise and self-generated sound into human body,” *Acta Paediatrica Japonica*, vol.32, no.6, pp.615-624 (1990)
- [36] 前川喜久雄, “脳科学への期待”, 音声研究, vol.8, no.3, pp.35-40 (2004)
- [37] フランシス・ハッペ, 自閉症の心の世界, 星和書店 (1997)
- [38] 渡部信一, 鉄腕アトムと真平君, ミネルヴァ書房 (1998)
- [39] J. F. Werker, “Who’s got rhythm?” *Science*, vol.288, pp.280-281 (2000)