実世界情報システムプロジェクト～視聴覚研究グループ～

構造不変の定理に基づく音声の構造的表象

峯松 信明

情報理工学系研究科 電子情報学専攻

概 要

音声コミュニケーションにおいて，音声の生成，収録，伝送，再生，聴取の何れの段階においても不可避的に音響歪みが混入する。声道長，調音器官の構造的差異を含む話者性，収録・伝送・再生音響機器の差異，収録環境の差異，更には聴取者間の聴覚特性の差異などが不可避的に混入する。音響音声学及び音声工学は「全ての音声は歪んでいる。歪んでいない音声を取得する唯一の手段は発声しないこと，聴取しないこと」と主張する。その一方で人間にとっては，歪みだらけの音声が最も「楽な」メディアである。さて，これら不可避的な音響歪みを表現する次元を保有しない音声の物理的表象が，言語学の一部である構造音韻論の物理実装を通して，峯松により提案されている（音響的普遍構造）[1, 2, 3]。本報告ではまず，この構造的表象の数学的解釈について述べる。その中で，本表象と相対論，量子論，情報論との接点について述べる。更に，構造的に表象された二つの音声事象間の距離尺度を解析的に導出する。

1 音声に不可避的に混入する歪み

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。一方，乗算性歪みや線形変換性歪みの消去は，概要に示したように「発声しないこと，聴取しないこと」を意味するため，これらは不可避である。

乗算性歪みはマイクや伝送特性などのフィルタリングであり，また，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に対するベクトルbの加算となる（$c' = c + b$）。

声道長差異によるフォルマントシフトは，スペクトルに対する周波数ウォーピングとして捉えられる。また，メル･バーク尺度は聴覚特性の平均パターンに過ぎず，各聴取者は異なる周波数ウォーピング関数を聴覚系に有する。この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 [4]（線形変換性歪み，$c' = Ac$）。

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（$b_1, b_2, , , A_1, A_2, , ,$）による統合歪みも一次変換となり，本報告ではこれを，不可避的な音響的歪みの工学的モデルとして採用する。

2 構造音韻論とゲシュタルト心理学

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」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。そこには**図 2** に示す音素群が成す幾何学構造の意識があり，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。音韻論では，音声に不可避な非言語的情報を頭の中で消失させ，彼らの考える言語の本質を議論する。この心的過程を数学的に明確化し，音声物理から，真の意味で言語的情報のみを抽出することを考える。

本研究で検討する物理表象は，単音を単位として音声物理を要素分割･個別観測するのではなく，音声事象群が成す構造・まとまり・調和を捉える操作に相当する。これは，人間の知覚が「対象の

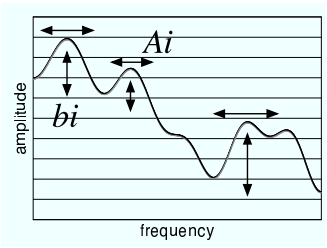

図 1. 不可避的な歪み

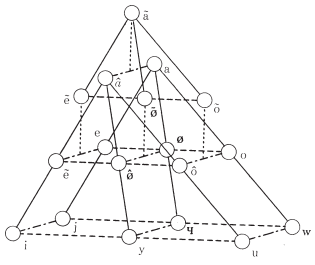

図 2. 構造音韻論

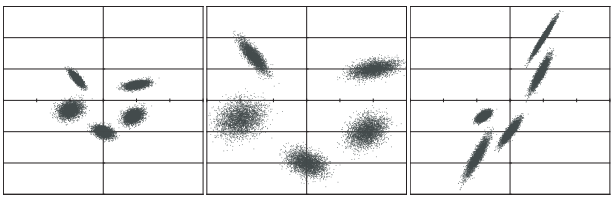

図 3. 構造不変の定理（これらが全て同一構造となる）

個別刺激を統合して起こるものではなく，それ以前に全体的な枠組の中で知覚が起こる」とするゲシュタルト心理学における音認知と同一視することもできる。従来，ゲシュタルト心理学と音認知との接点は聴覚情景分析において盛んに議論されてきたが，本研究ではこの考えを音声言語情報処理に導入するという側面もある。

ソシュールを発端とする構造主義においても，エーレンフェルスを発端とするゲシュタルト心理学においても，構造の普遍性・不変性，及び，その構造の変換性・移調性が等しく議論されている。本研究では音声物理に対して，普遍・不変の構造を仮定し，不可避な非言語的情報はその構造の変換・移調によって具現化されるとして議論を進める。

さて**第 1 節**に示したように，不可避的な非言語的情報は，ケプストラムで張られるユークリッド空間において一次変換，即ち，アフィン変換として記述される。空間内に，n 個の音韻を n 点で張られる構造として表現すると，この構造はアフィン変換により不可避的に歪んでくる。[4] において行列表現された変換には，回転（鏡像も含む）や遷移以外の変形要素が含まれるからである。即ち，数学的に音韻構造は，必ず話者・収録環境・聴取者間で歪むことになる。

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。以下，筆頭著者が「構造不変の定理」として定義している数学的事実を元に説明するが,「空間を歪ませて観測する」という方法論は，光速度を不変にするために空間・時間を再定義し，重力の発生を説明するために空間を歪ませて議論した相対性理論と，数学的には全く等質である。

3 構造不変の定理に基づく構造音韻論の物理的実装

空間内に存在する n 点に対して，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，二つの確率密度分布に対して算出される「∞ 次元空間における正規化相互相関」の $-\ln$ と等価であり [5]，即ち，二分布間の “関係” を “距離” として定義している。なお，距離単位であるが，$0 \leq \int_{-\infty}^{\infty} \sqrt{pq}dx \leq 1.0$ 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。

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

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，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。

$$BD = \frac{1}{8}\mu_{ij}^T \left(\frac{\sum_i + \sum_j}{2}\right)^{-1} \mu_{ij} + \frac{1}{2}\ln \frac{|(\sum_i + \sum_j)/2|}{|\sum_i|^{\frac{1}{2}}|\sum_j|^{\frac{1}{2}}}$$

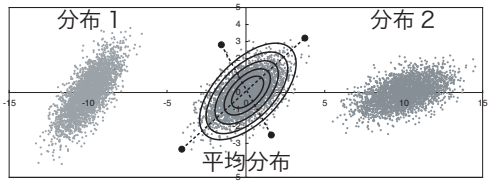

図 4. 平均分布形状とそれに基づく空間の歪み

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）。第一項であるが，これ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を仮定した場合の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（微分幾何学の一種）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，

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

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

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

とすると，計量 $M_{\mu\sigma}$ は以下のように算出される。

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

定数項を無視すると，これはフィッシャー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，計量としてフィッシャー計量を用いた場合の空間は，情報幾何学 [6] における多様体そのものである。即ち，構造音韻論の物理実装はリーマン幾何学が呈する歪んだ空間を必要とし，その空間は情報幾何学における多様体であり，音韻構造はその多様体内の n 点（即ち n 分布）として定義され，この情報論的構造は非言語情報によって一切不変となる（A=回転，b=遷移）。

さて，構造不変の定理は，物理事象を分布（混合ガウス分布）として表象することを要求する（ぼやけた n 分布）。その一方で，情報論的構造（距離行列）を抽出してしまうと，それは明確な n 角形へと変換され，変換後は元の物理現象がどのような分布形状をしていたのかは観測不可能になる。量子論における観測問題，即ち「観測することにより分布が点となる」現象に対して，フォン・ノイマンやペンローズらが「観測者（人間）の意識が分布を点へと縮退させる」との仮説を呈している。この仮説は，多くの理論物理学者によって否定されている。しかし，本数学定理を，音声事象群（ぼやけた n 分布群）が人間の聴取意識（或いは無意識的操作）によって明確な n 角形（即ち言語的情報・音韻構造）に変換されると解釈し，かつそれが，音声言語に対するゲシュタルトであると解釈した場合，彼らの仮説との類似点は非常に興味深い。しかも，得られる明確な構造は，情報幾何学が規定する構造，即ち “情報” である。

4 発声の構造化と超分節的特徴

既に筆者は，提案手法に基づく音声事象の構造化を外国語発音学習に応用し，実際に高校の英語教育において，一部試験的に導入している。ここでは 40 文ほどの発声より音素モデルを構築し（n 分布），構造抽出を行ない，非言語情報を表現する次元を失った音声表象として学習者一人一人を描いている（個人の構造化，発音カルテ）[7]。

しかし，構造不変の定理は，原理的に，音素や音節といった言語単位を要求しない。また複数の発声も要求しない。音声は常に揺れているように，一つの発声を考えた場合でも，各区間のスペクトル系列を分布として捉えれば，発声は分布系列に変換される。そして時間的に離れた事象間も含め，全ての二分布間距離を求めれば，構造を規定することになる。これを非言語情報を失った，純粋に言語情報のみを有する音声言語ゲシュタルトと定義する。音素や音節といった分節区間よりも長い

音声区間を対象として構造化をかけた場合，この構造は超分節的特徴として定義される。即ち本構造化手法は，音声の分節的特徴のみを用いて，音声の超分節的特徴を定義していることに他ならない（即ち，ゲシュタルト）。また文献 [8] で述べているように，構造のサイズは音声事象の継続時間や，強弱勢などを表現する。即ち提案する音声の物理表象は（現段階では）F_0 情報以外の，スペクトル，パワー，継続長などの情報を融合して形成される，音声の統合的物理表象となっている。

5 構造間距離尺度

二つの単語発声を構造的に表象し比較することで，二発声の言語的同一性の判定，即ち音声認識はできるのだろうか？可能な場合，音声認識は，各単音の物理特性（フォルマント周波数，スペクトル形状）を直接的には一切必要としない，ということになる。各音に対するピッチの絶対知覚無しに音楽が楽しめるように（音楽ゲシュタルト），各単音に対する音韻の絶対知覚無しに，その音声言語ゲシュタルトの知覚のみで音声を楽しむことは可能なのだろうか？

提案する音声の構造的表象の音声認識への応用可能性については文献 [9] を参照して戴きたい。本報ではその距離尺度について示す。なお紙面の都合上，導出の詳細については別途機会を設けることとし，ここでは得られた結果のみを示す。構造不変の定理により，構造の回転と遷移のみを考えればよい。そこで，ユークリッド空間における二つの M 点群（$P_1, ..., P_M, Q_1, ..., Q_M$）に対して，構造 Q を回転と遷移のみ（即ち，直交行列とベクトルによるアフィン変換）で構造 P に近づけ，対応する点間距離の和（$\sum_{i=1}^{M} \overline{P_iQ_i}^2$）の最小値を求める。その最小値は以下となる。

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

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 次正方行列 S^tTT^tS の固有値である。S は行列 $(O\vec{P}_1, ..., O\vec{P}_M)$ であり，T は行列 $(O\vec{Q}_1, ..., O\vec{Q}_M)$ である。

なお上記構造間距離尺度は，構造不変の定理が定義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するからである。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文献 [2] で示した通りである。

6 まとめ

相対論・量子論・情報論を融合させることで「構造不変の定理」を導出し，不可避的に混入する非言語情報を表現する次元を消失させることで，音声言語ゲシュタルトを定義した。この音声言語ゲシュタルトは構造音韻論の物理実装と解釈される。従来の音声科学・工学は韻律研究を除き，ほぼ全て，音声を要素分割（分節化）することで議論を繰り返してきた。本報告は，一連の音声研究に対するアンチテーゼであるが，要素分割型の音声研究との融合も，当然可能である。

参考文献

[1] 峯松，音講論，2-7-9，pp.281–282 (2004-3)

[2] N. Minematsu, Proc. ICASSP, pp.585–588 (2004-4)

[3] N. Minematsu, Proc. ICASSP, (2005-3, accepted)

[4] M. Pitz *et al.*, Proc. Eurospeech, pp.1445–1448 (2003)

[5] A. Bhattacharyya, Bull. Calcutta Math. Soc., vol.35, pp.99–109 (1943)

[6] 甘利他，“情報幾何学の方法”，岩波講座応用数学，岩波書店 (1993)

[7] 峯松，信学技法，TL2004-47，pp.47–52 (2004-12)

[8] 朝川他，信学技報，SP2004-28, pp.53–58 (2004-6)

[9] 丸山他，音講論，1-5-14，pp.27–28 (2005-3)

[10] N. Minematsu, Proc. ICSLP, pp.1669–1672 (2004-10)

[11] N. Minematsu, Proc. ICSLP, pp.1317–1320 (2004-10)

[12] 峯松他，信学技法，SP2004-27, pp.47–52 (2004-6)

[13] 藤野他，音講論，2-5-3，pp.59–60 (2005-3)

[14] 朝川他，音講論，2-1-11，pp.225–226 (2005-3)

[15] 村上他，音講論，2-P-9，pp.379–380 (2004-10)