

# 音声に内在する音響的普遍構造\*

峯松 信明 (東大・情報理工)

## 1 はじめに

音声の分節的側面に対する新しい物理表象を提案する。対数スペクトル包絡に対する一次変換によって算出されるパラメータ (例えばケプストラム) を用いて瞬時的な分節的特徴を記述する。パラメータ時系列となる音声事象を確率論的に有限個の状態 (分布) として記述し, 任意の状態間の距離を情報論的に算出し, 最終的に音声事象を, 状態群が成す構造として相対論的に捉える。こうして構造化された音声事象はパラメータに対する任意の一次変換  $Ax + b$  に対して不変となる。ベクトル  $b$  は構造のシフトとなり, 行列  $A$  は構造の回転となる。ケプストラムをパラメータとした場合, 乗算性及び線形変換性歪みは各々, 構造のシフトと回転になる。即ち提案する音声表象は, 乗算性及び線形変換性の歪みを表現する次元を理論的に保持し得ない音声の物理表象である。

## 2 言語音の記述と言語音構造の記述

### 2.1 音声学的世界観の限界

音声認識に代表される, 音声を受理する形態の系には sheep and goats が存在することが常識として扱われている。受理系構築の際に音及び言語のモデルが学習されるが, その学習環境と, 受理系の使用環境とに不一致が存在した場合, 系の性能は劣化する。音モデルの不一致問題の解決法として, モデルを使用環境に適応させる方法, 或いは, 入力パラメータにおける「使用環境らしさ」を正規化する方法が議論されている。しかし適応, 正規化の何れの場合も少量データでの遂行が必須であり, 処理精度の劣化を招き, その結果 goats を完全に抹消することは困難となる。

音声の分節的側面に混入する歪みとして, 加算性, 乗算性, 及び線形変換性の歪みを考える。ここで加算性歪みは「その雑音源を原理的に抹消できる」と言う意味において他の二つとは異なる。例えば話者性の一部は乗算性の歪みであり, 音声を収録, 伝送, 再生すれば, 音響機器による乗算性歪みは混入する。声道長の違いは周波数ウォーピングとして表現されるが, これは線形変換性歪みの一種である。またバーク尺度によって近似される聴覚特性も, その聴覚特性を変数変換によって線形軸上で考察すれば入力音の周波数ウォーピングとなる。つまり, 音声の生成・収録・伝送・再生・聴取の行為そのものが「歪み」であり, その意味において「歪んでいない音声」は存在しない。

何故 goats は消えないのだろうか? 全ての音声処理系は「音のモデル」を保持することが当然の如く考えられている。これは「個々の言語音を記述する」音声学の世界観に基づいている。しかし「個々の音」を記

述すれば, 乗算性・線形変換性の歪みもそのまま記述されることになる。即ち, 音声学的世界観で音声表象を求める限り goats を抹消することは不可能である。

### 2.2 音韻論的世界観による音声表象

適応・正規化処理の問題は利用可能なデータ量だけであろうか? そもそも, 適応・正規化処理を行なっても, 乗算性・線形変換性歪みを表現する次元は音モデルの中に依然として残っている。Goats の完全消滅を目指す場合, それは, これら歪みを表現する次元を保持しない音声表象によって初めて実現され则认为すべきである。本研究の目的は「乗算性・線形変換性の歪みを表現する次元を理論的に保持し得ない音声の物理表象」が存在することを示すことにある。

音声学的世界観からの脱却が求められている。ここでは音韻論的世界観にその答えを模索することを考える。以下その理由を示す。音韻論は, 音声物理に纏る種々の歪みを (人間の頭の中で) 抽象化というプロセスを経ることで抹消し, 言語音群に観測される種々の言語的現象を議論する科学である。そこでは「言語音の並び」に内在する規則・構造や, 「言語音群の中」に内在する規則・構造の明示化が対象となる。当然乗算性・線形変換性の歪みは存在しない。人間の頭による音声物理の抽象化の結末が音韻論における議論であるならば, その議論を物理の上で実装することにより, 音声学的世界観と音韻論的世界観の物理的差異が明らかとなる。そしてその差異 (物理に基づく抽象化) を経ることで, 音声学的世界観が「乗算性・線形変換性の歪みを表現する次元を理論的に保持し得ない音声の物理表象」へと変貌することが期待される。

音韻論的世界観に基づく言語音記述の一例として, Halle によるロシア語の音素樹型図を図 1 に示す。これは, 弁別素性と自然類に基づき, 対象言語に観測される言語的現象を考慮して行なわれる音素のトップダウンクラスタリングである [1]。以下, 音素群の構造化を通して目的とする音声の物理表象を追及する。

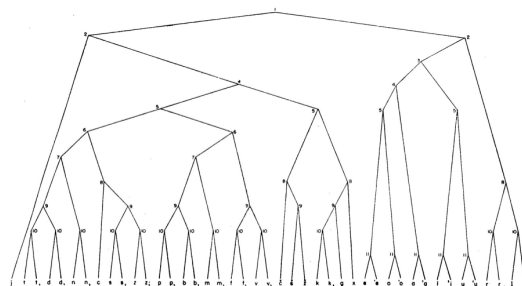


Fig. 1.1. Branching diagram representing the morphemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. oral vs. nonoral; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality; 6. strident vs. nonstrident; 7. nasal vs. nonnasal; 8. continuant vs. noncontinuant; 9. voiced vs. voiceless; 10. sharp vs. plain; 11. accented vs. unaccented. Left branches represent minus values, and right branches, plus values for the particular feature.

図 1. Halle によるロシア語音素の樹型図

\*The acoustic universal structure in speech

### 3 確率論・情報論に基づく音声事象の構造化

Halle による音素分類は、言語的現象との整合性を基に行なわれており、言語的現象の解釈の違いによって複数の分類が得られる。また、対象言語に対する言語的知識の蓄積がなければ分類そのものが不可能である。これらは当然、音声の物理的表象に望まれる性質ではない。本研究では、言語的現象との関連を敢えて切り離すことで音声の構造化（相対論的表象）を考える。つまり、ボトムアップクラスタリングである。

#### 3.1 確率論に基づく音声事象の状態（分布）化

パラメータ時系列となった音声事象を有限個の状態として記述する。これは音声認識で広く使われる隠れマルコフモデルにおける状態化で実現される。各状態は確率論的にパラメータの分布として記述される。

#### 3.2 情報論に基づく音声事象の構造化

状態間距離をバタチャリヤ距離を用いて算出する。

$$\begin{aligned} BD(u, v) &= -\ln \int_{-\infty}^{\infty} \sqrt{P_u(x)P_v(x)} dx \\ &= \frac{1}{8} \mu_{uv} \left( \frac{\sum_u + \sum_v}{2} \right)^{-1} \mu_{uv}^T + \frac{1}{2} \ln \frac{|\sum_u + \sum_v|/2}{|\sum_u|^{1/2} |\sum_v|^{1/2}} \end{aligned}$$

バタチャリヤ距離は、二つの確率密度に対して両事象の独立性を仮定した上で同時確率密度を求め、その平方根を全積分して同時確率へ変換し、その対数で定義される。即ち二つの事象の同時確率に対する自己情報量であり、情報論的に距離を定義していることになる。本距離尺度そのものはガウス性を要求しないが、両分布が単一ガウス分布であれば上式で算出される。

与えられた（ある空間内の） $n$  点に対して、 ${}_nC_2$  個存在する全対角線の長さを考慮することは（距離行列を求めることは）、 $n$  点で構成される構造の情報を考えることに等しい。例えば、全音素を状態として記述し、その状態間距離行列に対してクラスタリング分析を行なえば、図 1 のような音素樹型図が得られる。

#### 3.3 構造化された音声事象が持つ特徴

各状態を学習する際に使用された学習データ（ケプストラムベクトル  $c$ ）に対して、一次変換をかける。

$$c'_t = Ac_t + b$$

平均ベクトルと分散共分散行列は以下となる。

$$\begin{aligned} \mu' &= E(c'_t) = A\mu + b \\ \Sigma' &= E(c'_t - \mu')(c'_t - \mu')^T = A\Sigma A^T \end{aligned}$$

上記の変換式は、任意の共通の一次変換によってバタチャリヤ距離は不変であるという性質を導く。

$$\begin{aligned} BD(\mu'_u, \Sigma'_u, \mu'_v, \Sigma'_v) &= BD(A\mu_u + b, A\Sigma_u A^T, A\mu_v + b, A\Sigma_v A^T) \\ &= BD(\mu_u, \Sigma_u, \mu_v, \Sigma_v) \end{aligned}$$

これは、情報論的に算出された状態間距離は一次変換に対して不変であり、構造として表象された音声事象は一次変換に対して不変であることを意味する。

ユークリッド空間における構造（点の集合）に対する一次変換は、構造に対して種々の変形を施す。 $A$  は構造のスケーリング、せん断、回転及びそれらの組み合わせを、 $b$  は構造のシフトを意味する。ここで点が平均ベクトルであり、そこに分布が付随している状態を考える。上記した種々の変形の中で、変形前後で二点（分布）間距離が不変であるのは回転とシフトのみである。つまり、 $A$  を掛ける演算は構造の回転となり、 $b$  を足す演算は構造のシフトとして観測される。

第 2.1 節で述べたように、音声の生成・収録・伝送・再生・聴取の行為そのものが不可避免的に乗算性及び線形変換性の歪みを混入する。しかし、音声事象を確率論的に有限個の状態として記述し、情報論的に状態間距離を定義し、最終的に、音声事象を相対論的に構造として捉えたと、その構造はこれら歪みに対して一切影響を受けない。即ちこれが「乗算性・線形変換性の歪みを表現する次元を理論的に保持し得ない音声の物理表象」である。この構造を「発声者及び聴取者の生理的条件、及び収録・伝送・再生環境の音響的条件に依存せず普遍的に観測される」という意味において、音声の音響の普遍構造と定義する。既に自明であるが、構造として表象された音声事象は話し手から聞き手に至るまで（雑音源の抹消が可能な加算性歪みを無視すると）理論的に歪み得ない。つまり、話し手と聞き手との間に情報の損失・改変が理論的に起こり得ないコミュニケーションチャンネルが存在することになる。「このチャンネルはどのような情報を伝達できるのか」という議論は今後の課題であるが、従来の音声科学・工学の議論の中で、このような完全なチャンネルが存在することを示した例を筆者は知らない。

### 4 まとめ

確率論、情報論、そして相対論という三つのツールを通して音声事象を眺めた時に、半世紀以上前に行なわれた言語学の議論が純然たる物理に落ちることを示し、完全なるコミュニケーションチャンネルの存在を示した。既にこの新しい音声表象に基づく分析・応用に関する実験的検討を開始している [2, 3, 4, 5, 6]。

### 参考文献

- [1] M. Halle, The sound patterns of Russian, The Hague: Mouton (1959)
- [2] 峯松, “音声に内在する音響の普遍構造とそれに基づく語学学習者モデリング”, 電子情報通信学会音声研究会資料, SP2003-179 (2004)
- [3] 峯松, “音声の音響的普遍構造の歪みに着目した外国語発音の自動評定”, 電子情報通信学会音声研究会資料, SP2003-180 (2004)
- [4] 峯松, “音響的普遍構造と言語的普遍構造間の整合性に基づく発音明瞭度の評定”, 電子情報通信学会音声研究会資料, SP2003-181 (2004)
- [5] 朝川他, “音声の音響的普遍構造に着目した英語強勢・弱勢母音の分析”, 日本音響学会春季講演論文集 (2004)
- [6] 浜野他, “音声の音響的普遍構造に着目した非言語情報推定に関する実験的検討”, 日本音響学会春季講演論文集 (2004)