実世界情報システムプロジェクト～視聴覚研究グループ～

音声に内在する音響的普遍構造とそれに基づく音声情報処理

峯松 信明

情報理工学系研究科 電子情報学専攻

概 要

音声というメディアは常に歪んでいる。発声者が異なれば，発声者間の生理学的差異による歪みが混入し，収録・伝送・再生すれば音響機器の歪みが混入し，更には，聴取者が異なれば，聴取者間の聴覚特性の差異による歪みが混入する。歪みゼロの音声を得るための唯一の方法は「話さないこと・聴かないこと」である。歪みだらけの音声に対する音声情報処理の提供する方法論は「常時適応・正規化」である。話者が変わる，マイクが変わる，条件が一つ変わる毎にモデル適応，パラメータ正規化を行なう。その一方で，人は歪みだらけの音声を「最も “楽な” 意思伝達手段」と感じる。本研究は，計算機上に実装されている音声処理系と人間が感じる感覚とのズレを埋めるべく，音声コミュニケーションに隠されたある数学的な “からくり” を示す。と同時に，人間がその “からくり” を使用していること，及びその “からくり” に基づく音声アプリケーションの例を示す。

1 はじめに

音声コミュニケーションにおいて不可避的に混入する歪み（の源）は大きく，発声，収録・伝送・再生，及び聴取の 3 種類に分類される。なお，ここでは物理的消滅が可能な加算性雑音は考慮しない。発声者が異なれば，声道長サイズ（即ち体のサイズ）に基づく音響的差異が生じる。これは，対数スペクトルの周波数ウォーピングとして観測される。また，GMM による話者認識において長時間スペクトルの平均パターンが話者性を表すように，話者性の一部は伝達関数（即ちフィルター）として実装される。収録・伝送・再生の一連の操作は伝達関数を縦続にかける形での歪みを混入する。最終的に音声が聴取される時も，バーク尺度で近似される歪みが混入される。この場合，音声刺激が歪むのではなく，観測系の周波数軸が非線形に伸縮する。変数変換により周波数軸を線形にすると，音声刺激の方が歪むこととなるが，この歪みは周波数ウォーピングとなる。結局，3 種類の歪みは伝達関数をかける形の歪みと周波数ウォーピングという形の歪みの 2 つに分類されることとなり，最終的に音声コミュニケーションによってもたらされる歪みは，ケプストラムを c と書くと，$c' = Ac + b$ となり，一次変換として記述される。

対数スペクトルの系列として音声を表象する音声工学では，これら A と b は常に歪みとして観測され，観測される度に，適応処理或いは正規化処理を必要としている。しかし，人はそのような処理を自らが行なうという意識を持つことがない。本研究は，これらの不可避的な歪みを表現する次元を持たない音声の物理表象が存在することを示し，その表象に基づく情報処理について検討する。

2 音韻論の物理実装に基づく新しい音声表象

2.1 音韻論における言語音構造の明示化

ここでは，音韻論における音声表象に着目する。音韻論とは，音声の年齢，性別，話者性といった非言語情報を一切無視（抽象化）し，音声の中に含まれる純粋に言語的情報のみに着眼する。そして「音の並び」に内在する規則，関係，構造，あるいは「音群の中」に内在する規則，関係，構造を明示化する。非言語情報の抽象化は音韻論者の

頭の中で行なわれるが，その抽象化を行なった末の議論を物理の上で実装することができれば，乗算性及び線形変換性の歪みである非言語情報を表現する一切の次元を保有しない音声の物理表象が実現されると期待される。

言語音構造記述の一例として，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。言語的現象の解釈によって異なる構造が呈されることとなるが，これは，音声の物理表象としては望ましくない。本研究では，音構造と言語的現象との関連性を敢えて断ち切り，言語音群に内在する関係，構造を純粋にボトムアップ的に構築することを考える。

2.2 音韻論の物理実装に対する必要十分条件

与えられた n 個の要素群のボトムアップクラスタリングは，一般的に，任意の二要素間の距離のみの情報（距離行列）によって行なうことができる。空間内の n 点に対して，${}_nC_2$ 個だけ存在する線分の長さを規定することは，n 点で構成される構造を規定することと等しい。音韻論における議論は，この構造が，話者，収録環境に因らず（即ち $c' = Ac + b$ の変換に因らず），普遍的に観測されることを主張している。結局，音韻論の主張を物理実装するための必要十分条件は，

- 空間内に n 点で構成される構造（任意の二点間距離）が，アフィン変換で不変である。

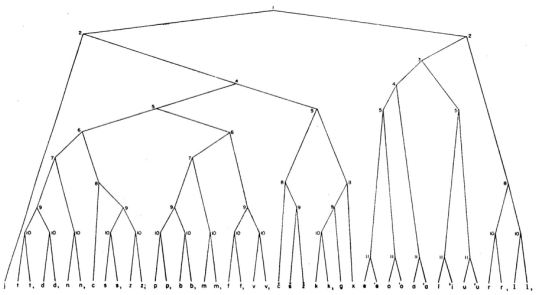

Fig. I-1. Branching diagram representing the morphonemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. consonantal vs. nonconsonantal; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality; 6. strident vs. mellow; 7. nasal vs. nonnasal; 8. continuant vs. interrupted; 9. voiced vs. voiceless; 10. sharped 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

となるが，これは数学的に不可能である。唯一の可能性は A が，回転や鏡像要素しか持たない行列となることであるが，先行研究によりその可能性も打ち消される。結局，個々の言語音をケプストラム空間内の一点で記述する方法論では，音韻論の議論の物理実装は数学的に不可能である。

2.3 情報理論に基づく音韻論の物理実装

ある特定話者によって発声された各言語音 P_i を（多次元）ガウス分布で近似することを考える。

$$P_i = \mathcal{N}(\mu_i, \Sigma_i) \tag{1}$$

この場合，アフィン変換 $c' = Ac + b$ によって μ, Σ は以下のように変化する。

$$\mu' = E(c') = A\mu + b \tag{2}$$

$$\Sigma' = E(c' - \mu')(c' - \mu')^T = A\Sigma A^T \tag{3}$$

結局，各言語音を分布として捉えた場合，音韻論の物理実装は，以下の条件を満たす空間（距離尺度）を選定する問題となる。

- 任意の二分布間距離がアフィン変換前後において不変である。

この条件を満たす分布間距離としてバタチャリヤ距離がある。

$$\begin{aligned} &BD(i,j) \\ &= -\ln \int_{-\infty}^{\infty} \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})}\, d\boldsymbol{x} \\ &= \frac{1}{8}\mu_{ij} \left(\frac{\sum_i + \sum_j}{2} \right)^{-1} \mu_{ij}^T + \frac{1}{2} \ln \frac{|(\sum_i + \sum_j)/2|}{|\sum_i|^{\frac{1}{2}} |\sum_j|^{\frac{1}{2}}} \end{aligned} \tag{4}$$

μ_i は i の平均ベクトル，μ_{ij} は $\mu_i - \mu_j$ を，Σ_i は i の分散共分散行列を意味する。なおバタチャリヤ距離は上式からも分かる様に，二つの確率密度，$p_i(\boldsymbol{x})$ と $p_j(\boldsymbol{x})$ 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，その平方根に対して全領域で積分する形で確率の次元へ変換し，その

対数をとることで（即ち自己情報量）距離を定義している。以下の等式が成立する。

$$
\begin{aligned}
&BD(\mu_i', \Sigma_i', \mu_j', \Sigma_j') \\
&= BD(A\mu_i + b, A\Sigma_i A^T, A\mu_j + b, A\Sigma_j A^T) \\
&= BD(\mu_i, \Sigma_i, \mu_j, \Sigma_j)
\end{aligned} \tag{5}
$$

上記の事実は，二話者・環境特性の差異がアフィン変換で記述されれば，両者によって発声された音声資料から得られる言語音群構造には一切差異が無いことを意味する。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有しない音声（言語音群）の物理表象」であり，言語学の一分野である音韻論で議論される言語音構造の物理実装が可能であることを意味する。この構造を以下，音声に内在する音響的普遍構造と呼ぶ。

3 種々の音声事象の構造化

3.1 言語の構造化から個人の構造化へ

ある個人が発声した音声サンプルから構造抽出することを考える。ある言語の母語話者であれば，どの話者を用いても凡そ同じ構造を呈する。しかし，外国語発音における構造は，たとえ母国語が同じであっても二話者間で異なることが容易に想像される。既に峯松により，外国語学習者の「今」を記述する発音カルテとしてこの構造抽出が用いられており，そこには，性別・年齢・話者性・収録機器特性・伝送特性と無縁な，主に外国語発音における母国語依存性のみが表現されている [2]。**図 2** に日本人学生一名による音声サンプルから生成した英語音素樹型図を示す。日本人特有の癖が随所に見られる。[3, 4] では，距離行列をベクトルとして見なして計算される行列間のユークリッド距離が，近似的に，乗算性・線形変換性歪みに関する適応・正規化処理を施した後の音響マッチング距離になることを示し，従来の音響マッチングに基づく発音評定ではおよそ不可能と思われる処理を，容易に実現している。

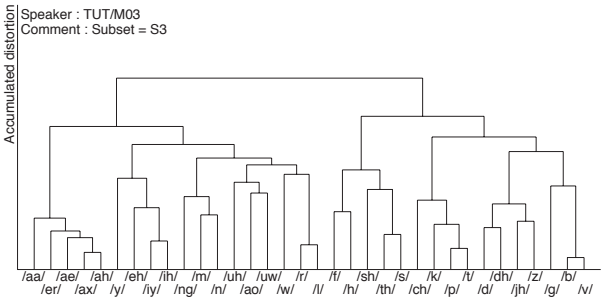

図 2. 日本人学生による英語音素の樹型図

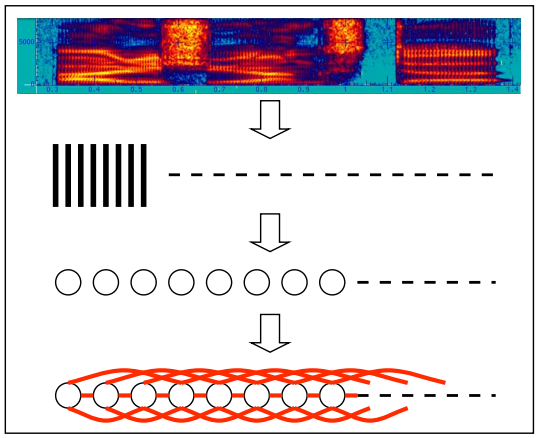

図 3. 発声の構造化

3.2 個人の構造化から発声の構造化へ

言語，個人の構造化はいずれも，音素という言語的に有意味な単位を音響的に分布として捉え，情報理論的に構造化をかけることで，不可避的な歪みの除去を実現した。しかし，歪みの除去は分布群の構造化によってもたらされるのであり，分布が言語的に意味のある単位を構成する必要はない。音声事象を分布群として近似することは，単発声の音声サンプルでも可能であり，そこから構造を構成することも可能である。**図 3** に発声単位での構造化について示す。単発声から抽出された構造としての情報は，音声コミュニケーションにおける発声・収録・伝送・再生・聴取という各段階で不可避的に混入する歪みに何ら影響を受けず，回転（A）と平行移動（b）をするだけで，話し手から聞き手に完全無欠のまま伝達される。

4 構造化された音声事象を用いたコミュニケーション

4.1 不特定話者音声

図 3 で示した発話の構造化は，時間的に離れた事象間の距離のみを使い，複数個の音響事象間で構造を構成する。この時，A や b が時不変であれば，異なる A や b の間で（即ち異なる話者間で）事象間距離は変わらない。HMM 合成技術を用いると任意の時点で話者性を（スペクトル的に滑らかに）変化させることが可能であり，この技術を用いると，本来とは異なる事象間距離が観測される刺激音声が得られる。この，不特定話者音声に対する人間の反応を見ることで，完全無欠のコミュニケーションチャネルが音声知覚過程において利用されているか否かについて実験的に検討する。

4.2 不特定話者音声聴取実験

話者性を変えるタイミング制御として，8 モーラ（話者性変化無し），4，2，1 モーラ，1 音素，1 状態の 6 段階を用いた。F_0 の制御に関しては外部から与えることとした。8 モーラ無意味モーラ列であるので，LHHHLLLL という F_0 パターンを与えた。被験者としては音声研究者（合成音の聴取実験に慣れた被験者）5 名と，初めて合成音声の聴取実験に臨む 3 名である。構造として音声を捉えるということは，音声ストリームをより広い範囲で捉える処理であり，また，構造ではなく個々の音響事象を捉えるということは，狭い範囲で音声を捉える処理となる。これは知覚単位という言葉で呼ばれる現象に相当するが，音声研究者の場合分析的な聴取が，非音声研究者の場合は非分析的な聴取が期待されるため，非音声研究者の結果に話者性変化のためのモーラ同定率劣化が観測されると予測される。また，不特定話者音響モデルによる音声認識性能についても検討する。この場合，音響事象間の関係を見るというプロセスは全く実装されていないため，話者性変化に因らず一定の同定率となることが予測される。結果を**図 4**，**図 5** に示す。予測した通り，非音声研究者は話者変化の頻度が高くなるにつれ同定率は下がる。この結果は，音声をより大きな単位で構造として捉える処理が存在することを指示する。なお，状態単位で話者性を変えると，HMM 合成のスムージングの結果，話者性が明瞭に表出されないため，同定率は跳ね上がっている。

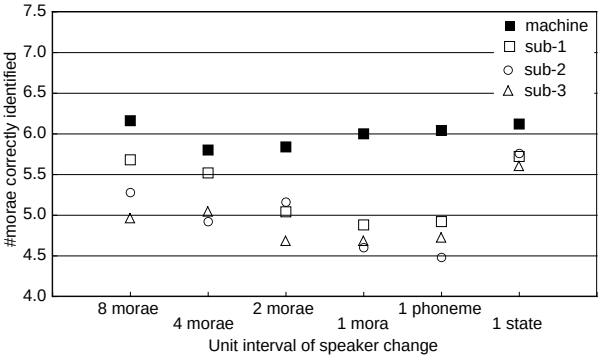

図 4. 非音声研究者によるモーラ同定率

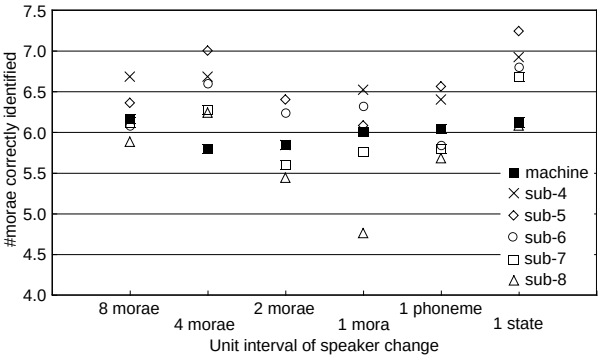

図 5. 音声研究者によるモーラ同定率

発表文献（一部）

[1] 峯松他，“音声に内在する音響的普遍構造とそれに基づく音声コミュニケーション”，話し言葉の科学と工学ワークショップ講演論文集 (2004)

[2] 峯松，“音声に内在する音響的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，SP2003-179，pp.25–30 (2004)

[3] 峯松，“音声の音響的普遍構造の歪みに着眼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，SP2003-180，pp.31–36 (2004)

[4] 峯松，“音響的普遍構造と言語的普遍構造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，SP2003-181，pp.37–42 (2004)

[5] N. Minematsu, “Yet another acoustic representation of speech sounds,” Proc. ICASSP’2004 (2004)