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

峯松　信明† 　松井　 健†† 　広瀬　啓吉†††

† 東京大学大学院情報理工学系研究科
†† 東京大学工学部
††† 東京大学大学院新領域創生科学研究科
〒 113–0031 東京都文京区本郷 7–3–1
E-mail: †{mine,matsuken,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　人間間の音声コミュニケーションでは，音声の生成，収録，伝送，再生，聴取の何れの過程においても，乗算性或いは線形変換性の歪みが不可避的に混入する。これらの歪みは，その発生源の物理的消滅が可能である加算性雑音とは異なり，原理的に消すことができない。音声認識で利用される音響モデルは，これらの歪みに対処するために，適応化・正規化処理を必要とすることが多い。さて，人が対話相手を変えた時に「少量の音声データによる，自らの耳の適応処理，或いは，相手の音声の正規化処理を行なっている」との意識を持つことがあるだろうか？本研究は，この「素朴な疑問」に対する一つの解を提供する。乗算性及び線形変換性歪みは，性別，年齢，話者性，収録機器・伝送特性などに相当するが，これら歪みを表現する次元を一切保有しない音声の物理的表象が存在することを数学的に示す。歪みを表現する次元を保有しないため，この物理的表象は話し手から聞き手に至るまで，何ら改変されることなく完全無欠に伝達されることとなる。話者性が時間軸上で変化する合成音声を用いた音声知覚実験により，人間が，この理論的に情報が歪み得ない音声コミュニケーションチャネルを利用していることを示す。

キーワード　音響的普遍構造，乗算性・線形変換性歪み，音韻論，音声コミュニケーション，知覚実験

The acoustic universal structure in speech and speech communication based on the structure

Nobuaki MINEMATSU†, Ken MATSUI††, and Keikichi HIROSE†††

† Graduate School of Information Science and Technology, University of Tokyo
†† Faculty of Engineering, University of Tokyo
††† Graduate School of Frontier Sciences, University of Tokyo,
7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–0031 Japan
E-mail: †{mine,matsuken,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Speech communication between humans has several steps such as production, encoding, transmission, decoding, and hearing. In every step, multiplicative or linear transformational distortions are inevitably involved. Unlike additive noise, source of which can be eliminated physically, deletion of sources of the above distortions always disables speech communication. In speech recognition, to deal with the distortions, model adaptation or parameter normalization is often required. The authors wonder whether a human listener always adapts his ears or normalizes input speech before he starts to understand what a new speaker says. This paper shows a possible answer to this simple question. It is mathematically proved that acoustic representation of speech exists where completely no dimensions are allowed for multiplicative and linear transformational distortions. Since the new representation cannot convey the above two kinds of distortions at all, a speech event represented by the proposed method can be transmitted from a speaker to a listener without any alterations. Results of perceptual experiments with synthetic speech samples of dynamically variable speaker individualities indicate that human listeners employ the theoretically distortion-free channel for efficient and reliable speech communication.

Key words　universal structure, inevitable distortions, phonology, speech communication, perceptual experiment

1. はじめに

近年の計算機技術の発展と数理統計的モデリング技術の融合によって，音声情報処理技術は飛躍的に進歩した。大語彙連続音声認識に代表される認識技術（HMM&N-gram）の高精度化，波形素片接続型音声合成に代表される合成技術の高精度化などがよい例である。これらの技術は共に，豊富なデータ量を武器に，ボトムアップ的に現象を分類し，必要に応じて汎用モデルによるパラメータ化を行ない，それらに基づいて音声を認識あるいは合成するパラダイムである。データ量が豊富にある場合は，高精度に動作するシステム構築が可能であるが，入手できるデータ量の増加が原理的に難しい場合，処理系の高精度化は困難になる。また，対象とする現象が多岐に渡る場合は，データ収集に奔走することとなる。HMM&N-gram に代表される音声認識の場合，子供，高齢者，非母語話者などの特殊話者の音声，更には感情音声などはデータ収集そのものが困難であり，その結果，データ量を「当て」にした方法論では困難な面がある。その場合，適応・正規化技術を用いて，モデルパラメータや入力パラメータを変形することが頻繁に行なわれる。しかし「適応・正規化用データ量をできるだけ少なくしたい」という要望が生じたり，語学学習などのアプリケーションでは，学習者に適応するとスコアが上昇するため「適応技術の導入そのものが困難」といった状況に遭遇することもある。一方，波形素片接続型の音声合成の場合，例えば感情音声合成を考えると，当然感情音声のデータベースが必要となる。感情の種類，度合いまでを考慮すると，必要なデータ量は増加の一途を辿る。何れにせよ，ボトムアップ的な解決方法は，問題の高度化と共に，その解決手段の困難さが増大してしまう。

豊富なデータ量に基づくボトムアップ的なアプローチとは対照的に，対象とする現象の原理・原則・からくりを解き明かし，その知見に基づいた専用モデルを構築することで問題解決を図る方法論も存在する。例えば基本周波数（以降，F_0）パターンのモデリングとして広く知られる生成過程モデル [1] は，声帯振動を司る筋の動きをモデル化することで，F_0 パターンを近似する数式を導いている。このモデルは非常に少量のパラメータで F_0 パターンを精度良く近似することが可能である。汎用モデルの場合は，データからそれを生じさせた要因（即ちパラメータ値）を推定するアルゴリズムが完備していることが多いが，専用モデルの場合は必ずしも整備されている訳ではない。生成過程モデルでも，実測の F_0 パターンからのパラメータ自動推定は，読み上げ音声の場合でも，決して容易ではない [2]。

本研究では，音声コミュニケーションに対して，数学的に示される一つの「からくり」について報告する。この「からくり」に基づく音声の物理表象では，性別，年齢，話者性，収録環境特性，伝送特性など，音声コミュニケーションに不可避に混入する静的な歪みを示す次元を一切保有しない。即ち，本表象によって表現された音声は，これら歪みによって一切歪まず，完全無欠のまま聞き手に伝達されることになる。本研究では更に，音声知覚実験を通して，人間がこのコミュニケーションチャネルを活用して音声を受理している様子を示す。

2. 音声認識技術に対する素朴な疑問

本研究では，今では常識である「不特定話者音響モデル＋話者・環境適応（及びパラメータ正規化）」という音声認識における音響モデリングの常套手段に「素朴な疑問」を投げ掛ける。

- 音声活動の本質が対話である場合，人が聞く音声の半分は「自らの声」である。このような状況で，人の耳に不特定話者音響モデルに相当する「もの」が構築されるのだろうか？人の耳に宿るのは大部分が「自らの声」モデルではないのか？
- 人は自らの耳を話者や環境に対して，常に適応しながら音声を処理しているのだろうか？意識の中で「今，適応中」という感覚を覚えたことは，少なくとも筆者らにはあまり無い。

当然，機械による音声認識処理系と人間による音声認識処理系が，同一のものである必要は無い。しかし，上記のような「素朴な疑問」を考えた場合，音声の物理の中に，まだ音声研究者が気付いていない「からくり」があることを予期させる。例えば Moore は，人間一人が一生を通して聞く音声量を学習データとして，現在の方法論で音声認識システムを構築しても，それは人間の性能には遠く及ばないことを予測している [3]。

音声コミュニケーションでは，音声の生成，収録，伝送，再生，聴取の何れの過程においても不可避的に乗算性或いは線形変換性の歪みが混入する。その歪みが学習時には観測されないタイプの歪みであった場合，音響モデルと入力音声間の音響的不一致が生じ，システム性能は劣化する。これらの歪みは，雑音源の物理的な消滅が可能な加算性雑音とは本質的に異なり，消滅が不可能な歪みである。人が発声すれば，その話者の声道長に基づく歪み（線形変換性の歪みの一種 [4]）が混入し，また，話者性の一部は乗算性の歪みとなる。発声された音声を収録，伝送，再生すれば当然乗算性の歪みが混入する。更には，聴取の段階においても歪みが混入する。バーク尺度で知られる聴覚特性（入力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。変数変換により軸を線形軸とすれば，入力刺激の歪みとなり，その歪みは線形変換性歪みの一種となる [4]）も個人差は存在し，そのための歪みが入る。つまり，生成，収録，伝送，再生，聴取の行為そのものが歪みであって，その意味で，歪んでいない音声は存在しない。歪みゼロの音声をもたらす唯一の行為は「発声しないこと，聴取しないこと」である。

人間は音声というメディアを用いて何不自由無くコミュニケーションを遂行している。恐らく最も心的な処理タスク量が少ないメディアが音声であろう。しかし音声は，原理的に「歪みゼロ」という実体が存在し得ない，歪みだらけのメディアである。この事実に対する現在の音声認識技術の回答は「常時適応・正規化」である。しかし，筆者らはその処理を自らの中に意識したことがない。現在の音声認識技術に対して問い掛けたい「素朴な疑問」とは，結局のところ次の一節に集約される。

- 不可避的に混入する乗算性・線形変換性歪みを表現する次元を保有しない音声の物理的表象が存在し，それに基づくコミュニケーションチャネルが存在するのではないのか？

3. 音韻論の物理実装に基づく新しい音声表象

3.1 個々の音を記述する方法論の限界

音声工学は，音響音声学を基盤としている。音響音声学は，言語の種類を問わず，言語音の一つ一つをその音響的特性に基づいて記述することを目的とした科学である。調音音声学は，各音をその調音的特性（調音位置，調音様式）を用いて記述する科学である。調音音声学に基づく調音形態の一つを異なる二話者が実装した場合，当然，その音響特性は両者の間で異なる。即ち，調音音声学は，音声の音響的特性を抽象化した記述方式を用いている科学であると言える。逆に言えば，音響音声学では，調音音声学において無視している差異までも見てしまっている。現在の音声工学では，この音響音声学が提供する「文字」に基づいて各音を記述し，その統計モデルでもって最終的な「音」モデルとすることが多い。大量のデータから統計モデルを作ることで，話者特性・環境特性が消えることが期待されるが(注1)，例えば不特定話者音響モデルを用いたとしても，システムが不得手なユーザが存在するのは事実である。たとえ適応・正規化処理を行なったとしても，それは話者や環境特性を “1” に近付けただけであり，完全に “1” になる訳ではない。その結果，適応化の効果も当然，ユーザ依存となる。これらの事実は「不可避な歪みを隠さずに記述する文字による表象」を「沢山集めること」の限界を示唆している。例えば，教育や医学などシステムの安定性がまず第一に求められるような応用分野では，方法論的な不完全性に基づく不安定性は忌み嫌われる。この不完全性を完全に断ち切るためには「沢山集めること」や少量データで「表象を修正すること」ではなく，本来は「不可避的な歪みを表現する次元を保有しない文字を見つけること」によって初めて実現される，と考えるべきであろう。

3.2 個々の音が織りなす系を記述する方法論

そのような都合のよい物理的表象が存在するのだろうか？そもそも，音声の物理現象から，性別，年齢，話者性，収録機器・伝送特性などの不可避な歪みを表現する次元を消失させることなど，可能なのだろうか？音声学と対を成して音声科学を構成する科学として音韻論がある。本研究では，音韻論における言語音の表象に着眼する。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（研究者の頭の中で）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議論の対象とならない。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,「個々の音の実体を記述する」音声学的世界観と「音の群に内在する構造を記述する」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。

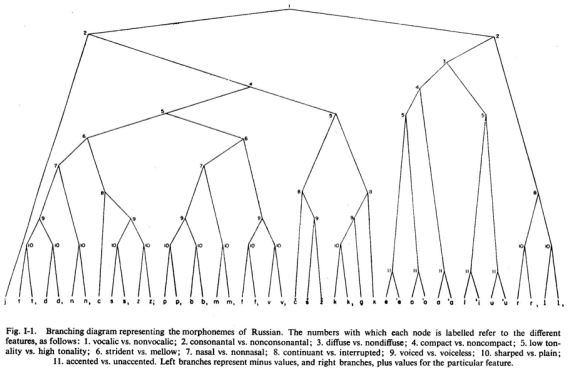

Fig. I-1. Branching diagram representing the morphonemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. consonantal vs. nonconsonantal; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality; 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
Fig. 1　Halle’s phoneme diagram of Russian

音韻論における「言語音群の中」に内在する規則・構造の明示化の一例として，Halle によるロシア語の音素樹型図を**図 1** に示す [5]。この樹型図は，弁別素性を用いて構成される。一つの素性によって音素セットは二分され，二つの素性によって四つに分類される。どのようにして素性を選択するか，であるが，最終的に得られた樹型図の各ノードに存在する音素サブセットが，そのサブセットに特有の言語現象を有するように素性を選ぶ（この場合，その音素サブセットは自然類を成す，と言われる）。弁別素性，及び，素性と言語的現象に着眼した音素分類は Jakobson によって提案されたが [6]，その根底には，Saussure の構造主義の哲学がある [7]。

Halle による音素分類は，triphone 学習における状態共有時に行なわれるクラスタリングと同じように，全音素を一端一つにまとめ上げ，それを何らかの知識に基づいて適切な二分割を繰り返すトップダウンクラスタリングである。両者の違いは，二分割の選択を音響的な妥当性に求めるか，言語的な妥当性のみに求めるか，である。言語的な妥当性のみに基づいた場合，当然，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，対象言語に対する知識の蓄積が無ければ分類そのものが不可能である。言語音群の中に内在する規則の物理実装を考えた場合，これらは望まれる性質ではない。本研究では，言語的現象や言語知識との関連を敢えて切り離すことで，言語音群の構造化を考える。即ち，ボトムアップクラスタリングである。

3.3 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，収録環境に因らず普遍的に観測されることを主張している。この主張を以下，数学的に考察する。

ある特定話者によって発声された個々の言語音が，ケプストラム空間内(注2)において，n 個の「点」として存在していると

(注1)：これらの要因をノイズ，即ち平均値がゼロ，だと考えれば，大量のデータを集めることで話者特性・環境特性は消えることが期待される。

(注2)：最終的には，対数スペクトル包絡に対する一次変換で規定される係数で

する。この各点は，話者の違い，収録環境の違いによって動くことになる。**第 2. 節**で述べたように，声道長（の個人差），あるいは，聴覚特性（の個人差）はいずれも，周波数ウォーピングという形で歪みを生み，これは，ケプストラムドメインでは，行列 A を掛ける演算として数学的に導かれる [4]。一方，話者性の一部や，収録機器特性，伝送特性は伝達関数を掛ける演算となり，これはケプストラムドメインではベクトル b を足す演算となる。結局，話者の違い，収録環境の違いは，一次変換（アフィン変換）によって近似されることになり，点 c は $c' = Ac + b$ へと変化する。以上の考察より，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。

ということになるが，これは数学的に不可能である。唯一の可能性は，声道長，聴覚特性を近似する 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, [4], [8] によれば，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（言い換えれば，各言語音を一枚のスペクトルスライスで代表させる方法論）では，音韻論の議論の物理実装は数学的に不可能である。

3.4 情報理論に基づく音韻論の物理実装

各言語音を一枚のスペクトルスライスで代表させる方法論では，音韻論の主張は数学的に不可能である（話者性，収録環境特性の違いによって構造が変化してしまう）ことが示された。次に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（即ち自己情報量を求めることで）距離を定義している。即ち「二つの事象 i と j が同時に起こる」という事象に対して，その自己情報量を用い，情報理論に基づいた距離の定量化をしている。バタチャリヤ距離尺度そのものは両分布がガウス分布に従うことを要求せず，両分布が単一ガウス分布である場合のバタチャリヤ距離の計算式が上式である。上述した様に，この距離尺度の下では以下の等式が成立する。

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

異なる二話者が異なる環境で発声した音声資料を用いて，話者別に，各音素モデルをガウス分布としてモデル化し，音素群構造（距離行列）をバタチャリヤ距離尺度を用いて構成することを考える。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，両者の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に対する，完全なる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。

上記の議論は，分布間距離がアフィン変換前後で変化しない距離尺度であればバタチャリヤ距離である必要は無い。当然仮定する分布形によって採択する距離尺度が変わることが予想される。バタチャリヤ距離の場合も，混合分布となったガウス分布に対して不変性が保たれるのか否かという問いもまだ未検討である。最終的には，音声の物理特性に最も適合する分布形と距離尺度を選択する必要があるだろう。

アフィン変換前後で分布間距離が不変である，という性質の幾何学的考察を行なう。アフィン変換による構造変化は，拡大・縮小，せん断，回転，鏡像，平行移動などに分類される。これらの中で任意の二点間の距離を変えない変換は，回転，鏡像，平行移動である。空間の次元を一つ増やすことで鏡像変換は，軸周りや平面周りの回転となることを考えると，二点間距離を変えない変換は，回転と平行移動に帰着される。つまり 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平行移動として解釈される。声道長の違いを人間の成長と考えれば，個人が呈する音響空間内における言語音構造は，長い年月を経ながら回転していることになる。そしてこの回転は聴覚特性によっても，もたらされると解釈される。

音声事象をスペクトル包絡（ケプストラムベクトル）の分布として記述する音声工学の言語音モデルから，本研究で提案する言語音群の構造（関係）モデルへの遷移を物理的抽象化という観点から考察する。二つの音の関係に対して平行移動を認めるということは，原点位置の情報を抽象化することに相当する。更に，二つの音の関係に対して回転を認めるということは，その関係が（実測時に）持っていた方向の情報を抽象化することに相当する。原点を捨て去ることで乗算性歪みが抹消され，方

張られる空間であればよい。

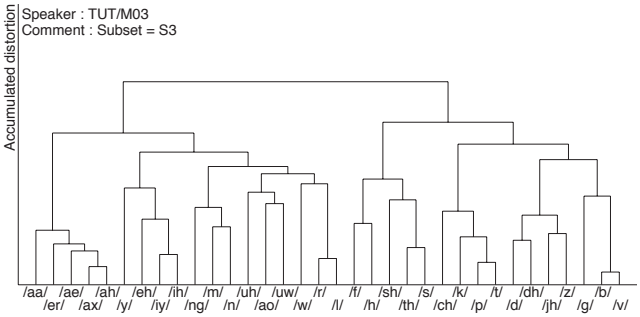

図 2　日本人学生による英語音素の樹型図

Fig. 2　English phoneme diagram produced by a Japanese student

向を捨て去ることで線形変換性歪みが抹消される。一見膨大な情報量を捨てているように見える。しかし，この音構造モデルだけに基づいた音声アプリケーションが可能であれば，そのアプリケーションの安定性は飛躍的に向上することが予想される。既に峯松はこの音構造モデルだけに基づいた音声表象を使って，語学教育応用のアプリケーションを検討している [9]〜[11]。

4.　種々の音声事象の構造化

4.1　言語の構造化から個人の構造化へ

音声事象から，不可避的に混入する種々の静的歪みを「そぎ落とした」形で定義される言語音群の物理的・構造的表象が存在することが明らかとなった。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こととなる（音韻論そのものである）。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 [9]。**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ていることが分かる。更に，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 [10], [11]。興味のある読者は是非参照して戴きたい。

4.2　個人の構造化から発声の構造化へ

第 3.4 節で述べた音韻論の物理実装（言語の構造化），及び前節で述べた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。パラメータベクトル時系列となった発声を一端，状態（分布）系列へと変換する。これは，HMM を単発声から学習すれば容易に実現される。その後，任意の状態間距離を計算し，距離行列を作成する。と同時に，各状態を記述する絶対的な特徴パラメータ（ケプストラム係数）を捨てる。こうすることで単発声は構造へと変換される。単発声から抽出された構造としての情報は，**第 2. 節**で示したように，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，回転と平行移動をするだけで，話し手から聞き手に完全無欠のまま伝達される。即ち，筆者らの素朴な疑問である「不可避的に混入する乗算性・線形変換性歪みを表現する次元を保有しない音声の物理的表象」が数学的に存在することが実証された。この「からくり」が本当に人間の脳内処理において実装されているのか，という問いに今すぐ答えることはできない。しかし，この「からくり」の存在可能性を意識した上で従来の音声知覚研究を眺めると，その存在を示唆していると考えられる種々の実験結果が得られていることに気付く。

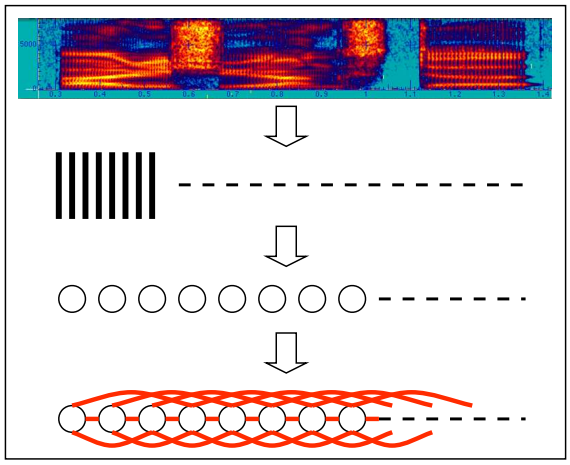

図 3　発声の構造化

Fig. 3　Structured utterance

例えば [12] では，連続音声中から切り出された母音や CV 音節を提示すると正しく同定されなくなる事実を報告している。既に述べてきたように，言語音の物理的実体は音声コミュニケーションチャネルの中で如何様にも変貌する。その一方で，分布として抽出された音響事象間（時間的に連続する事象のみならず，時間的に離れた事象間も含む）の距離は一切変化しない。連続音声中においてこのような普遍的とも言える特徴量を用いた処理を人間（脳）がしていると考えれば，切り出し音の聴取が困難になることは十分に頷ける。

また [13], [14] では，知覚単位のサイズとその処理について実験を行なっている。知覚単位として種々のサイズを想定し，より大きなサイズの知覚単位が利用できる環境では，より劣悪な（例えば SN 比が低い）音声でもその同定が可能であることを示している。この知見は，**図 3** における連続する有限個の音響事象群の構造化範囲が知覚単位であると仮定すると素直に理解できる。知覚単位の中に含まれる音響事象（**図 3** で言えば状態）数が n であるとすると，関係の数は ${}_nC_2$ となり，n の増加に伴

い，普遍的特徴の数は $O(n^2)$ で増える。即ち，より大きな範囲で音声ストリームを構造として捉えることができれば，入力音声が劣化していても単語の同定は容易に可能となる。

更に [15] は，日本語モーラ音声を使った知覚実験により，スペクトルの瞬時遷移量（絶対値）が最も大きい音声区間がモーラの知覚に最も貢献していることを示している。瞬時遷移量は連続する音響事象間の距離の一次近似であることを考えると，この実験結果は本研究で数学的に示した普遍特徴の利用を最も端的に示唆する実験結果であると言える。しかし，**図 3** に模擬的に示したように，音響事象間の距離は，音響事象の連続性を要求するものではない。時間的に離れた箇所の音響事象間の距離も利用した知覚過程の存在が示唆される。[15] の実験事実に基づき，スペクトルの動的特徴量（Δ ケプストラム）が提案された [16]。その後，スペクトルの遷移の様子をベクトルパラメータとして導入する数多くの試みが行なわれている。しかし，本研究で示した「からくり」を前提として考えると，スペクトル遷移を “ベクトル” として導入すると，連続する音響事象間の移動に関して “方向” という情報が入り，その結果，線形変換性の歪みに対する頑健性は低くなることが懸念される（但し実験的裏付けは無い）。なお，バタチャリヤ距離尺度は，周波数軸のメル尺度化に対しても興味深い性質を示す。メル尺度化はしばしば周波数ウォーピングとして実装されるが，これはケプストラムドメインでは A を掛ける演算となる。その結果，メル尺度化前後で分布間距離は一切変わらない。平均ベクトルのケプストラム係数値はメル尺度化によって変わるが，分散共分散行列も同様に変動し，結果として距離不変となる。HMM による音響マッチングでは，基本的に分布（モデル）と点（ケプストラムベクトル）の距離を求めており，この場合はメル尺度化によって距離は変化する。

以上，本研究で示した音声に内在する音響的普遍構造を既知として従来の音声知覚実験を考察し，また，現在広く用いられている音響モデリング技術についても考察した。次節では上記と異なる実験の枠組みを用いて，音声に内在する普遍特徴を用いて人間が音声を受理している様子を示す。と同時に，普遍的な構造の利用が実装されていない，機械による音声認識処理と，人間の音声知覚処理との差異についても実験的に考察する。

5. 不特定話者音声に対する音声知覚と音声認識

5.1 実験の目的

文献 [15] では，刺激音声の一部（前半或いは後半）を物理的に消すことで，スペクトルの瞬時遷移量の最大値（連続する二音響事象間の距離の近似値）が観測される音声区間の提示・非提示を制御している。本研究では音声を一切消さずに，本来存在すべき遷移量（音響事象間距離）とは異なる遷移量が存在する刺激音声を作成する。人間が発声した音声であれば，年齢・性別・個人性を問わず等しい遷移量が観測されるため，自然音声を用いて刺激音声の作成を行なうことは不可能である。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，この技術を用いることで，本来とは異なる遷移量が観測される刺激音声が得られる。

表 1　HMM 学習条件
Table 1　Conditions of training HMMs

話者・文セット	男性アナウンサー 7 名 / ATR503 文
サンプリング	16kHz / 16bit
窓	ハミング窓，25msec
フレーム周期	5 msec
パラメータ	メルケプストラム（0〜24 次元）
HMM	7 状態 5 分布，triphone

以下，話者性を任意の時点で変化させた合成音声を不特定話者音声と呼ぶ。本実験の目的は音響事象間の距離が正常及び異常な音声に対する人間及び機械（不特定話者音声認識）の 認識精度を見ることで，音声知覚過程において，普遍特徴がどのように活用されているのかについて検討することを目的とする。

5.2 刺激音声の作成

「無意味モーラ列の書き取り」をタスクとして選んだ。短期記憶の容量などを考慮し，系列長は 8 とした。男性アナウンサー 7 名による ATR503 文を用いて合成用の HMM を話者毎に学習した [17]。学習条件を**表 1** に示す。HMM 合成では通常，種々の言語情報を音素コンテキストとして用い，状態のトップダウンクラスタリング時に，音素環境・言語環境を統合的にマージする。本実験では，合成音の最終的な韻律制御は明示的に与える必要があるため，言語ラベル無しの HMM を作成した。

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 モーラ，2 モーラ，1 モーラ，1 音素，1 分布の 6 段階を用いた。なお，7 状態 5 分布の HMM であるので，分布単位で話者を変えるということは，1 音素当り 5 人の話者が登場することになる。合成用の HMM では，F_0，パワー，継続長など，音声合成に必要な全ての韻律情報もモデル化される。これらの情報をそのまま用いると話者性変換時の自然性が劣化する恐れがある。本研究では F_0 の制御に関してのみ明示的に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。具体的には，生成する各音素の継続時間長，及び，用いる話者群における平均 F_0 値を考慮し，生成過程モデルに基づいて F_0 パターンを生成した。継続長，パワーに関しては HMM に組み込まれた値をそのまま使用した。但し，分布単位で話者性を変える場合のみ，ある特定話者の時間情報を参照して音声を合成した。

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，及びモーラ認識の難度を考慮し，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。

5.3 実験手順と被験者

聴取実験は web 上で行なわれた。被験者はヘッドホンを通して 150 個の合成音声を聴取する。クリックにより提示が始まる。被験者は合成音声の提示と同時に，PC 上で書き取り作業を始める。聴取は 2 回まで許可した。なお，8 モーラ系列であること，及び一部音素（モーラ）は用いられていないことは事前に伝え，書き取りも 8 モーラで回答するよう指示した。

被験者としては正常な聴力を持つ成人男性 8 名であるが，内 5 名は音声研究に従事するものであり，聴取実験，或いは合成音声の聴取に比較的慣れた被験者である。残りの 3 名は今回初めて合成音声の聴取実験に望む被験者である。

聴取実験の後，提示した 150 種類の合成音声全てを，音声認識器により，連続モーラ認識した。音響モデルとしては CSRC 提供の性別非依存不特定話者音響モデルを，言語モデルとしては入力モーラ数が 8 であるとの制約を加味した認識文法を，デコーダとしては HVite v3.2.1(HTK) を用いた。

5.4 実験結果に対する予測

第 4.2 節で示した普遍的特徴，及び筆者らの一部による先行研究 [13], [14] 結果に基づいて，本実験より得られる結果を予測する。人間が普遍的な音響事象間の距離を積極的に利用しながら音声を受理していると仮定すると，話者性変化の間隔が短いほど（異常な音響事象間距離が生成される頻度が高いほど）モーラ同定率は下がるはずである。「音響事象間の距離を利用する」ということは，音声を比較的広い単位で受理する形態の知覚過程が優先的に働く状態を意味する。逆に言えば，音声を短い単位で受理する形態の知覚過程が優先的に働く状態で聴取すれば（即ち分析的な聴取），音響事象間距離の異常性に依らず（話者性変化の間隔に依らず）一定したモーラ同定率が示されることになる。被験者として音声研究従事者とそれ以外の 2 カテゴリの被験者を用いた理由はここにある。即ち，モーラ同定率が話者性変化によって劣化するとすれば，それは非音声研究従事者に見られる可能性が高い。さて，本実験では分布単位でも話者性を変化させている。音素中に 5 人の話者が登場する音声である。合成時に一分布から生成される音声長は平均約 15[msec] ほどであり，非常に短い。更に，HMM 合成は連続するフレーム間を滑らかに繋ぐ性質を持っており，その意味において個々の話者性が明確に表現されないまま音声が合成される可能性が高い。これらを考慮すると，話者性変化の頻度を上げていくとモーラ同定率は落ちるが，頻度を上げすぎると，逆に同定率は上がることが予測される。

一方，機械によるモーラ認識に関しては，音響事象間の距離を参照する枠組みは一切実装されていないため，話者性変化頻度に依らず，安定したモーラ同定率が示されるはずである。

5.5 結果と考察

図 4 に非音声研究者（sub-1〜sub-3）の結果を，**図 5** に音声研究者（sub-4〜sub-8）の結果を示す。8 モーラを単位とした場合の正解モーラ数の平均をプロットしている。自動認識の結果は両者において示している（■）。明らかなように，2 カテゴリの被験者グループは，モーラ同定率において絶対的な差異がある。非研究者の場合は自動認識よりも常に低いが，研究者の場合は一部を除いて自動認識よりも高くなっている。

まず非研究者であるが，前節で予想した通り，話者性変化の頻度が高くなるとモーラ同定率が低くなる傾向がある。しかし，話者性変化を分布単位にすると，同定率は急上昇する。これについても前節で考察した通り，話者性変化があまりに細かすぎる場合，HMM 合成の平滑化により話者性変化が十分に表現されない状況になるためであると考察される。一部 8 モーラ時

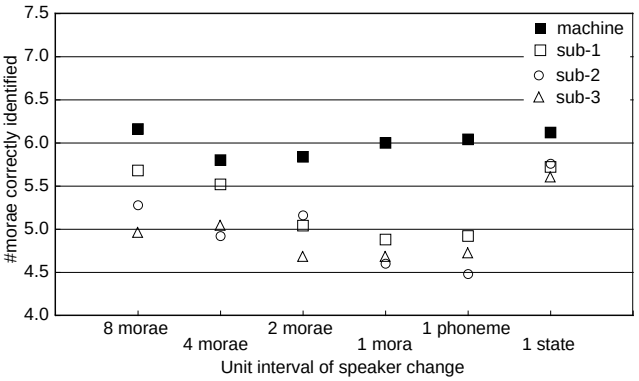

図 4　非音声研究者によるモーラ同定率

Fig. 4　Mora identification rates by naive listeners

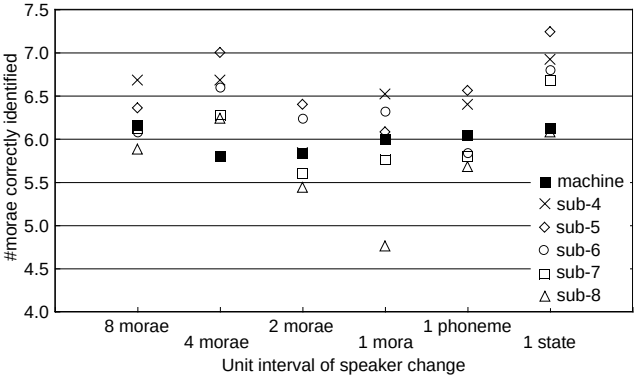

図 5　音声研究者によるモーラ同定率

Fig. 5　Mora identification rates by expert listeners

（話者性変化無し）の性能よりも高くなっている。これは分布単位で話者性を変えた合成音では，音響事象間距離がほぼ「本来の値」を示すことを意味するが，これについては現在調査中である。異なる話者変化間隔（但し分布単位を除く）における同定率の有意差検定を分散分析により行なったところ，以下の場合において危険率 10%未満となる差異（話者変化間隔がより短い方が同定率がより低い）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，p は音素を表す。以上のように，非研究者被験者の実験結果より，音響事象間の距離が異常値をとると，モーラの同定率が有意に減少することが示された。なお，自動認識の場合は任意の話者変化間隔間で有意な差は観測されなかった。

一方研究者の結果であるが，一部を除いて自動認識よりモーラ同定率が高い。また，話者変化間隔が短くなるに従いモーラ同定率が下がる傾向は一部の被験者（sub-8）にしか見られない。前節で予測したように研究者の場合，聴取が分析的であり，個々の音を捉える形での同定処理が優先的に行なわれていると考察される。しかし，分布単位で話者を変化させると同定率が高くなる傾向はここでも観測されている。上記と同様に話者変化の間隔が短すぎるために，変化の様子が十分に合成音の中に反映されないためであろう。なお sub-8 において，10%未満の有意差は 8m–1m($p = 1.79\%$)，4m–2m($p = 5.67\%$)，4m–1m($p = 0.35\%$) において観測された。

以上の結果を総合すると，聴取態度が過度に分析的で無い場合，人は音響事象間距離として存在する普遍的な特徴を一つのキーとして，より大きな時間範囲で音声ストリームを処理していることが示唆される。即ち，筆者らが投げ掛けた「素朴な疑問」である「不可避的に混入する乗算性・線形変換性歪みを表現する次元を保有しない音声の物理的表象に基づくコミュニケーションチャネル」の存在が実験的に示唆される。

6. ま と め

本研究ではまず，現在の音声認識技術に対し，

- 不可避的に混入する乗算性・線形変換性歪みを表現する次元を保有しない音声の物理的表象が存在し，それに基づくコミュニケーションチャネルが存在するのではないのか？

という「素朴な疑問」を投げ掛け，音韻論における言語音表象の物理実装を通して，乗算性・線形変換性歪みを表現する次元を理論的に保有しない，音声の物理表象が存在することを数学的に示した。即ち音声の音響事象を分布として捉え，その分布群を情報理論的に構造として捉えると，その構造は乗算性歪みに対しては平行移動し，線形変換性歪みに対しては回転することになる。次にその物理表象を用いて，発声者個人の構造化，更には単発声の構造化が可能であることを示した。単発声の構造化は，話し手の全ての発話が，乗算性・線形変換性歪みによって何ら改変せず，完全無欠のまま聞き手に伝わることを示唆する（但し，相対化・構造化されるので情報の曖昧性が生じることになる）。人間がこの完全なるコミュニケーションチャネルを利用しているのか否かに関して，話者性が時間軸上で随時変化する合成音声を用いて知覚実験的に検討を行なった。その結果，聴取態度が過度に分析的で無い場合，人は音響事象間距離として存在する普遍的な特徴を一つのキーとして，より大きな時間範囲で音声ストリームを処理していることが示唆された。言い換えれば，人間にとって音声によるコミュニケーションが「楽である」という感覚は，このチャネルの存在の果たす役割が大きいことが推測される。今後，異なる形態の知覚実験を通してこのコミュニケーションチャネルの性質を明らかにすると共に，計算機上への実装を通して現在の音声認識技術との融合を検討する予定である。

文　　献

[1] H. Fujisaki *et al.*, “A model for synthesis of pitch contours of connected speech,” Annual Report, Engineering Research Institute, University of Tokyo, no.28, pp.53–60 (1969)

[2] 成澤修一，“音声の基本周波数パターン生成過程モデルの特徴パラメータ自動抽出法とそれを用いたコーパスベース韻律生成” 東京大学大学院情報理工学系研究科電子情報学専攻博士論文 (2003)

[3] R. K. Moore, “A comparison of data requirements of automatic speech recognition systems and human listeners,” Proc. EUROSPEECH, pp.2581–2584 (2003)

[4] M. Pitz *et al.*, “Vocal tract normalization as linear transformation of MFCC,” Proc. EUROSPEECH, pp.1445–1448 (2003)

[5] M. Halle, The sound patterns of Russian, The Hague: Mouton (1959)

[6] R. Jakobson *et al.*, Preliminaries to speech analysis: the distinctive features and their correlates, MIT Press, Cambridge (1952)

[7] S. E. Blache, The acquisition of distinctive features, University Park Press, Baltimore (1978)

[8] 江森正他，“音声認識のための高速最尤推定を用いた声道長正規化”，電子情報通信学会論文誌，vol.J83-D-II，no.11，pp.2108–2117 (2000)

[9] 峯松信明，“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資料，SP2003-179 (2004)

[10] 峯松信明，“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資料，SP2003-180 (2004)

[11] 峯松信明，“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)

[12] H. Kuwabara *et al.*, “Perception of vowels and CV syllables segmented from connected speech,” J. Acoust. Soc. Japan, 28, pp.225 (1972)

[13] 峯松信明，“人間における音声言語処理過程の分析とモデル化”，東京大学工学部電気工学科卒業論文 (1990)

[14] H. Fujisaki *et al.*, “Influence of context and knowledge on the perception of continuous speech,” Proc. ICSLP’90, pp.417–420 (1990)

[15] S. Furui, “On the role of spectral transition for speech perception,” J. Acoust. Soc. Am. 80 (4), pp.1016–1025 (1986)

[16] S. Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. Acoust., Speech, & Signal Processing, vol.34, pp.52–59 (1986)

[17] 徳田恵一，“HMM による音声合成の基礎”，電子情報通信学会音声研究会資料，SP2000-74，pp.43–50 (2000)