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

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　音声事象の物理実体を捨象し，全てを他者との関係のみで捉える音声の音響的表象を提案する。各音声事象を確率論的に分布として捉え，分布間の距離を情報論的に計算し，事象群が成す構造のみを相対論的に記述する。構造として表象された音声は，不可避的に混入する乗算性及び線形変換性歪み（話者・音響機器・聴覚特性の差異）によって一切歪まない。即ち音響現象が言語現象へと抽象されることを数学的に示す。単一発声を構造化すると，その構造は話し手から聞き手に至るまで，一切歪むことなく伝搬される。この構造を音響的普遍構造と呼ぶ。人間が生成する音声には普遍構造が常に存在するが，この構造が存在しない音声の合成が可能である。本研究では，普遍構造が存在しない合成音声を用いた知覚実験を通して，音声コミュニケーションにおける普遍構造の役割について検討する。
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

Yet another acoustic representation of speech based on physical implementation of structural phonology

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

† Graduate School of Information Science and Technology, 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　This paper proposes yet another acoustic representation of speech, where speech events are probabilistically described as distributions, distance between two of them is calculated based on information theory, the distributions are relatively captured as structure. The resulting structure is mathematically shown to have no dimensions to indicate multiplicative or linear transformational distortions. Once an utterance is structuralized, the structure can be transmitted from a speaker to a listener without any distortions. This is called acoustic universal structure in speech. This structure always exists in human speech but not always in synthetic speech. This paper also investigates the function of the structure through listening tests using the synthetic speech samples.
Key words　universal structure, inevitable distortions, phonology, speech communication, perceptual experiment

1. はじめに

音声コミュニケーションは，音声の生成，収録，伝送，再生，聴取という過程により構成されるが，何れの過程においても不可避的な歪みが混入する。話者の差異は声道形状の差異を生み，収録・伝送・再生時の音響機器の差異は機器特有の音響歪みを生み，聴取者の差異は聴覚特性の差異を生む。音声の物理表象として音響音声学に基づいたスペクトル表現が用いられてきたが，上記の要因によりこの物理表象は常に歪みを保有する。換言すれば，音響音声学及びそれに基づく音声工学は「音声は常に歪んでいる。歪みゼロの音声を取得する唯一の方法論は，発声しないこと，聴取しないこと，である」と主張する。

その一方で，人間はこの「常に歪んでいる」音声メディアを，「最も容易な」コミュニケーションメディアであると感じる。何故なのだろうか？音声工学が示してきた回答は「まずは沢山の話者，環境の音声を集めること」そして「常時適応・正規化すること」である。不特定話者モデルと呼ばれる形態の音響モデルを用いても当然網羅できない話者は存在し，結局，話者・発話スタイル・マイク・環境が変る度に適応・正規化と「忙しい」処理系を構築することでこの問題の解決を試みてきた。

問題は解けたのか？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準備したとしても，現在の方法論では人間の性能を示す音声認識システムの構築は困難であることを予測している [1]。「不可避的な歪み（静的な非言語

情報）が変る度にそれに追従する」方法論以外に解決手段は無いのだろうか？そもそも，音響言語変換には全く不要の非言語情報が直接見えてしまう物理表象を何故使うのか？音響音声学はある意味，音声音響学である(注1)。スペクトルは音の全てを表現する。この上で議論を展開すれば，常に「不必要なもの」への対処が生じる。「常時適応・正規化」はその良い例である。

本研究では，不可避的に混入する静的な非言語情報を表現する次元を保有しない音声の物理表象を数学的に導出する。本表象では「適応・正規化」とは完全に異なり，非言語情報を表現する次元そのものが抹消される。この物理表象は構造音韻論における古典的な議論を物理実装することで得られる。音韻論では静的な非言語情報に起因する，単音の音響差異・変動を全て研究者の頭の中で抽象化し（音素），音声の言語的側面だけに着眼する。即ち彼らの議論を物理空間において実装することができれば，それは「音響から言語への抽象化」の物理実装に繋がると期待される。そして，その抽象化の物理実装は可能である。

本研究は，音声コミュニケーションに潜むある「からくり」を示すが，この「からくり」を通して音声技術を眺めると，例えば，メルケプストラムやデルタケプストラムといった常套手段が，如何に「摩訶不思議な」代物であるか，が見えてくる。

2. 不可避的に混入する非言語情報のモデル化

提案する物理表象を述べる前に，不可避的に混入する非言語情報を工学的にモデル化する。そして，その工学モデルの上で，これらの次元を消す形で構造音韻論の物理実装を検討する。

音声認識の世界では，音声コミュニケーションに付随する雑音・歪みは主に三種類に分類される。加算性雑音，乗算性歪み，そして線形変換性歪みである。テレビ，ラジオなどの背景雑音が加算性雑音の典型例であるが，この種の雑音は考慮しない。それらは物理的な抹消が可能であり（スイッチを消す，別の部屋に移動するなど），不可避な雑音ではないからである。頭による抽象化以前に，体を動かすことで解消可能な雑音である。

収録系，伝送系，再生系の音響機器による差異が生じる歪みが乗算性歪みの典型例である。スペクトル系列の時間平均パターンが話者特性を近似できることからも分かるように，話者性の一部も乗算性歪みとなる。この種の歪みは意味論的にゼロ点が存在せず，抹消することは不可能である。音声事象をケプストラム(注2)ベクトル c で表現した場合，乗算性歪みはベクトル b の加算として表現され，その結果，$c' = c + b$ となる。

話者が異なれば声道形状は異なる。聴取者が異なれば聴覚特性は異なる。聴覚特性の差異は聴取後の聴覚イメージを聴取者間で異なるものにする。なお，メル尺度やバーク尺度は平均特性に過ぎない。これらは線形変換性歪みの典型例であり，不可避な歪みである。声道形状の差異，聴覚特性の差異も対数スペクトルに対する周波数ウォーピングとして工学的に実装される。[2] によれば，単調増加かつ連続であれば，如何なる周波数ウォーピングもケプストラムの次元では，行列 A を掛ける演算に数学的に変換される。即ち，$c' = Ac$ である。

音声コミュニケーションには様々な雑音・歪みが混入するが，不可避的なもののみに限定した場合，それは乗算性及び線形変換性の歪みとなる。異なる歪み源は異なる A 或いは b を呈するが（A_i と b_i），これらの歪みが統合されて得られる（統合）歪みも，最終的には $c' = Ac + b$ として記述されることになる。即ち不可避な歪みはアフィン変換としてモデル化される。

3. 音韻論の物理実装に基づく新しい音声表象

3.1 構造音韻論に基づく音声表象

音声の物理実体から非言語情報をそぎ落とすことで，音韻論者は何を見たのか？音韻論では，「音素の並び」或いは「音素群の中」に内在する規則・構造を論じる。ここでは後者の構造論に着眼する。Jakobson はフランス語の音素群を立体的に構造化し [3]，Halle はロシア語の音素群を樹型図として構造化した [4]。これらは弁別素性を用い，言語学的に妥当な構造を持つよう行なわれたトップダウンクラスタリング(注3)である。構造音韻論は「言語が含むのは，言語体系に先立って存在する観念でも音でもなくて，ただ，この体系から生じる概念的差異と音的差異とだけである」とする Saussure 言語学を発端とする [5]。

3.2 構造音韻論の物理実装に対する数学的な必要十分条件

トップダウンクラスタリングでは，対象物への知識の差異が異なる結果を導く。本研究では音韻論と異なり，ボトムアップクラスタリング(注4)による音素群の構造化を考える。この場合必要となるのは，n 個の要素（例えば音素）に対する任意の 2 要素間距離，即ち距離行列である(注5)。n 点に対する距離行列の取得は，${}_nC_2$ 個だけ存在する対角線の長さを取得することに等しい。そして対角線の長さを規定することは，構造を唯一に規定することに他ならない。即ちボトムアップクラスタリング（距離行列の視覚化）は，n 点の情報を，その点群が成す構造だけに着眼し，それを視覚化することと数学的に同値である。

n 点をその構造だけに着眼すれば，種々の情報がそぎ落とされる。このプロセスによって非言語情報の次元が一切抹消されれば構造音韻論の物理実装は可能となる。**第 2. 節**での議論を考慮すると，構造音韻論の物理実装に必要な必要十分条件は，

- 空間内に n 点で構成される構造（任意の二点間距離，距離行列）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。結局，個々の音素をケプストラム空間内の一点で記述する方法論（各音素を一枚のスペクトルスライスで代表させる方法論。各音素を一枚の MRI 調音画像で代表させる方法論も同様）では，構造音韻論（音韻構造）の物理実装は数学的に不可能である。

(注1)：更に言えば，単音音響学である。phonetics=phone-tics なのだから。

(注2)：スペクトル包絡間の差異を効率良く計算するために考案されたスペクトル包絡の表象。音声波形をフーリエ変換して得られるスペクトルを再度（逆）フーリエ変換し，その低次の係数のみを抽出したもの。スペクトル包絡を見ることと，ケプストラム係数を見ることは数学的には同値である。

(注3)：全ての要素を一端一つにまとめ上げ，対象物に関する知識を用いて 2 つ，4 つ，8 つと順次分類していく方法論。

(注4)：最も類似している 2 要素を統合し，要素数を順次減らしていき，最終的に 1 要素になるまでまとめあげる方法論。

(注5)：「言語には音的差異しかない」という Saussure の訴えをそのまま数学という言語に翻訳しただけである。

3.3 情報理論に基づく構造音韻論の物理実装

構造音韻論は幻影なのか？単一のピッチ波形を厳密に繰り返しただけではブザーにしかならないように，音声は常に揺れている。点では表現できないと考えるのが妥当である。ここでは，各音素 P_i を多次元ガウス分布 $\mathcal{N}(\mu_i, \Sigma_i)$ で近似する。アフィン変換 $c' = Ac + b$ によって μ，Σ は，各々，$\mu' = A\mu + b$，$\Sigma' = A\Sigma A^T$ へと変化する。結局，構造音韻論の物理実装は，以下の条件を満たす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。

$$BD(p_i(\boldsymbol{x}), p_j(\boldsymbol{x})) = -\ln \int_{-\infty}^{\infty} \sqrt{p_i(\boldsymbol{x}) p_j(\boldsymbol{x})} d\boldsymbol{x} \qquad (1)$$

二つの確率密度 $p_i(\boldsymbol{x})$ と $p_j(\boldsymbol{x})$ の相乗平均を(注6)，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。なお，全積分によって得られる値は「二つの分布 i，j からの二つのランダムサンプル群が同一の分布を形成する確率」として解釈できる。

なお，両分布が単一の多次元ガウス分布の場合以下となる。

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

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。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計算すると，二構造間の差異がアフィン変換である限り，両者の構造には一切差異が無い。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の物理表象」であり，この構造を以下，音声に内在する音響的普遍構造と呼ぶ。

点群によって構成される構造に対するアフィン変換は，幾何学的には，拡大・縮小，せん断，回転，平行移動などに分類される。これらの中で二点間距離が不変な変換は回転と平行移動である。即ち 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。人間の成長による音声の音響的変化が声道長の延びのみによって生じるならば，その変化は音韻構造の回転として解釈される。15 年ほどかけてゆっくり回る回転である。なお，音素を混合ガウス分布として近似した場合でも，その二分布間距離は一次変換によって不変である。

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

ある個人が発声した音声サンプルから音韻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（読み上げ文，方言，発話スタイルに差異がなければ）凡そ同じ構造が示される。しかし外国語発音の場合，二学習者間で異なる構造となることが容易に想像される。学習者の音韻構造を

(注6)：平方根をとらないと $BD(p_i, p_i) \equiv 0.0$ が満たされない。また一次変換に対する不変性も満たされない。

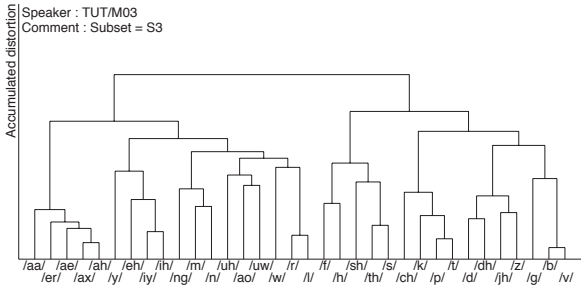

図 1　日本人学生による英語音素の樹型図

Fig. 1　English phoneme diagram produced by a Japanese student

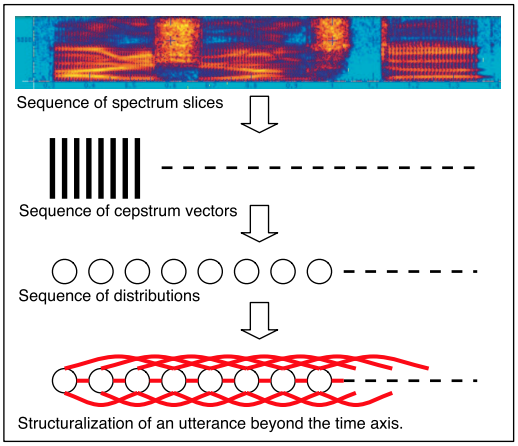

図 2　発声の構造化

Fig. 2　Utterance-level structuralization of speech

抽出すれば，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 [6]，学習者の「今」を記述する発音カルテとして利用できる。一例として**図 1** 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。なお，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実験的に示している。この距離尺度を用いて学習者間距離をスカラー量として定義し，学習者群の分類も行なっている。技術的観点から言えば，地球上に存在する 20 億人とも言われる英語学習者を彼ら独自の音韻構造に基づいて分類することも可能である [7]。

4.2 個人の構造化から発声の構造化へ

音響的普遍構造は，各音声事象を確率論的に分布として捉え，情報論的に分布間距離を計算し，分布群を構造として解釈することで定義される。ここで，音声事象が言語的に意味がある必要はなく，音響事象であっても成立する。そして，音声（音響）事象の分布化は単一の音声発声においても可能である。

図 2 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 の単発声のみによる学習そのものである。その後，任意の分布間距離を（時間軸を飛び越す形で）計算し，距離行列を作成する。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は，発声・収録・伝送・再生・聴

取において不可避的に混入する歪みに何ら影響を受けず，話し手から聞き手（の脳）に完全無欠のまま伝達される。この「からくり」を意識した上で従来の音声知覚研究を眺めると，その存在を示唆する種々の実験結果が得られていることに気付く。

[8] 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。言語音の物理的実体は環境によって変貌するが，音響事象間の距離は不変である。音響的普遍構造を活用しながら音声を捉えていると考えれば，切り出し音の聴取が困難になることは当然である。

[9], [10] では，知覚単位のサイズとその処理について実験している。より大きな知覚単位が利用できる場合，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 2** における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（分布）数が n の時，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。

[11] 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。しかし**図 2** に示したように，音響事象間の距離は，二音響事象の時間連続性を要求するものではない。入力音声を一端バッファリングし，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。

[11] の実験事実より，スペクトルの動的特徴量（Δ ケプストラム）が提案され [12]，その後，種々の動的特徴量が検討されている。しかし，本研究で示した「からくり」によれば，二音響事象間の “方向” の情報は，A を掛けることで変化（回転）する。即ちデルタケプストラムの導入によって，声道長の差異に対するロバスト性は必ず低下することが数学的に予測される。新たなパラメータの導入はデータ依存性を増す。動的特徴量の場合は（例えば）声道長への依存度が数学的に増すことになる。

周波数軸のメル尺度化も不思議な操作である。声道長の差異は各発声者に合わせようと努力するものの，聴覚特性の差異は平均パターンで代表させるだけである。両差異とも周波数ウォーピングとして表現される以上，音響的普遍構造においては一切歪みを生じない。周波数分解能を低域で上げ，高域で下げることでより効率的な符号化を可能とする変換技術との位置付けが妥当であり，本来認識精度とは関係の無い操作である。

時系列として存在する音響事象に対し，時間軸を飛び越える形で関係（距離）を捉えると，話者・環境・聴取者に依らない普遍的な物理構造が観測される（究極の抽象化）。この物理実体を言語的な単位（例えば音節）と定義すれば，音節系列に対し時間軸を越える形で関係を捉えたものが単語であり，単語系列に対してその関係（文法）を捉えたものが句，その上に，文，文脈，談話構造と，言語の階層構造を物理の上に再帰的に定義することも可能である（**図 3** 参照）。究極の抽象化とも言える

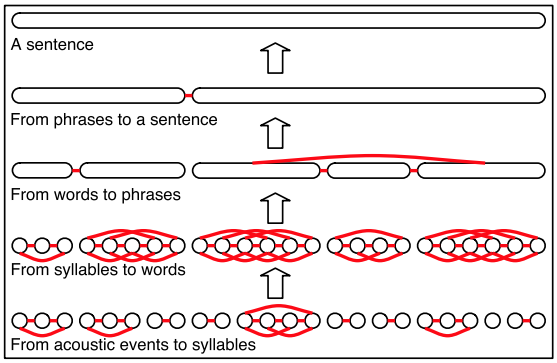

図 3　音響から言語へ

Fig. 3　From acoustics to linguistics

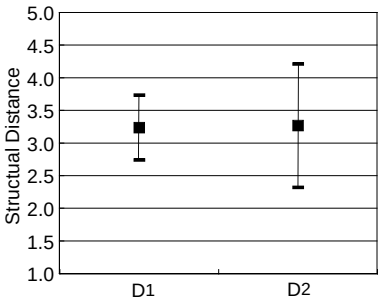

図 4　不可避的な非言語情報のそぎ落とし

Fig. 4　Cancellation of the inevitable distortions from speech

音響事象の構造化がヒトのみが持つ能力であれば，本研究で提案する音響的普遍構造を，例えばチョムスキーが主張する普遍文法 [13] の根源として捉えることはできないだろうか？

4.3　非言語情報のそぎ落とし

音響的普遍構造の抽出が非言語情報をどの程度そぎ落とすのかを示すために簡単な実験を行なった。日本人成人男女 2 名ずつによる孤立五母音発声を各話者 3 回ずつ収録した。12 個の普遍構造が得られるが，任意の 2 つの構造間差異を計測し，話者内差異（D_1）と話者間差異（D_2）で比較した。非言語情報がそぎ落とされていれば，話者間差異と話者内差異に有意差は無いはずである。構造間差異としては 2 つの距離行列 $\{P_{ij}\}$, $\{Q_{ij}\}$ に対する下式で定義する。2 つの構造を A 及び b に関して適応した後の，対応する音素間距離平均の近似値である [7]。

$$D = \sqrt{\frac{1}{M^2}\sum_{i<j}(P_{ij} - Q_{ij})^2} \qquad (3)$$

M は母音数，i, j は母音種類である。結果を**図 4** に示す。構造間差異は，話者内，話者間において有意差が無いことが分かる。なお，最小の構造間差異は話者間で観測されれている。

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

人間である以上，その発声は構造化によって全く歪むことなく聞き手に伝わる。音響的普遍構造が定義できない音声は自然界には存在しないと言える。しかし，音声合成技術を使えば，普遍構造不在の音声が可能である。HMM 合成技術を用いれば，任意の時点で話者性を（スペクトル的には滑らかに）変化させることが可能であり，この音声に観測される構造は当然歪んで

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

くる。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（不特定話者音響モデル）の認識精度を見ることで，音声知覚において，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。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 [9], [10] 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，離れた音響事象間の距離を参照する枠組みは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 5 に非音声研究者（sub-1〜sub-3）の結果を，**図 6** 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

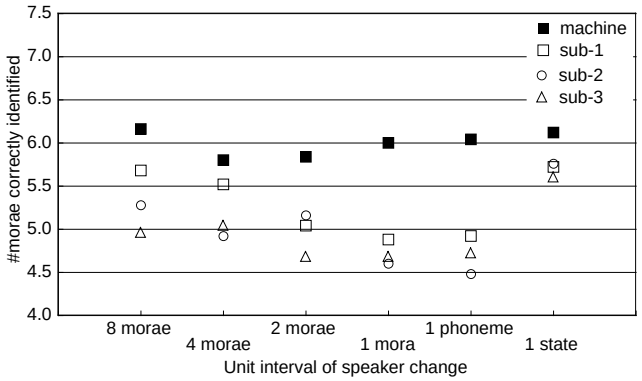

図 5　非音声研究者によるモーラ同定率

Fig. 5　Mora identification rates by naive listeners

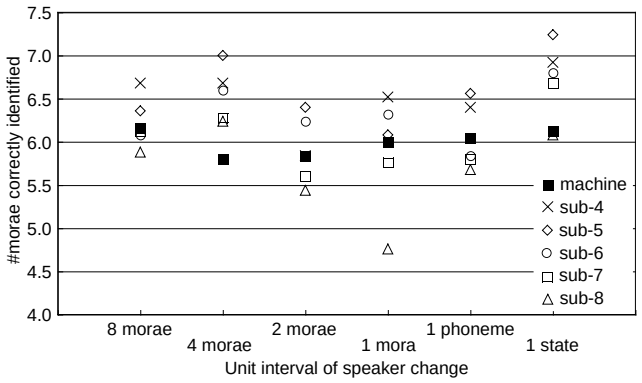

図 6　音声研究者によるモーラ同定率

Fig. 6　Mora identification rates by expert listeners

ラ時（話者性変化無し）の性能よりも高くなっている。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（過度に注意力を払わない楽な聴取を行なう）場合，人は音響事象間距離，及びそれにより構成される構造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することが示唆される。その一方で，分析的な態度で臨めば，関係の不自然さを無視した個々の音の同定も可能である。従来の音声認識は当然後者の聴取モードに相当する。現在，前者の枠組に基づく音声認識の検討を開始している。

6. ま と め

音声コミュニケーションの各過程において，種々の不可避的歪みが混入する。本研究ではこれら歪みを表現する次元を一切保有しない新しい音声の物理表象（音響的普遍構造）を，構造音韻論を物理実装する形で提案した。そして，静的な非言語情報を完全にそぎ落とす様子を実験的に示すと共に，人間が，この普遍構造に基づくコミュニケーションチャネルを利用していることを示唆する知覚実験結果を得た。音声コミュニケーションが「楽である」という感覚は，このチャネルの果たす役割が大きいと推測される。音響音声学及び従来の音声工学は「音声は常に歪んでいる。歪みゼロの音声を取得する唯一の方法論は，発声しないこと，聴取しないこと，である」と主張する。物理に落ちた音韻論は「音声の物理実体を捨象することで」定義されるもう一つの音声工学を予感させる。その工学は「音声は人間が発声する限り，数学的に歪むことが許されない。唯一歪ませる方法はパラ言語情報を付与することである [14]」と主張する。この二つの音声の物理表象及び工学は排他的なものではなく，協調的に音声を捉える形で位置づけることが可能である。

文　献

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

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

[3] ローマン・ヤコブソン，服部四郎編，“構造的音韻論”，岩波書店 (1996)

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

[5] ソシュール，小林英夫訳，“一般言語学講義”，岩波書店 (2003)

[6] 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”，信学技報，SP2003-179 (2004)

[7] 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”，信学技報，SP2003-180 (2004)

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

[9] 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)

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

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

[12] S. Furui, “Speaker independent isolated word recognition using dynamic features of speech spectrum,” IEEE Trans. ASSP, vol.34, pp.52–59 (1986)

[13] 酒井邦嘉，言語の脳科学，中公新書 (2002)

[14] 浜野他，“音声の分節的特徴に着眼したパラ・非言語情報推定に関する実験的検討”，信学技報, pp.143–150 (2004)