社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

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

峯松　信明†

† 東京大学大学院情報理工学系研究科
〒 113–0031 東京都文京区本郷 7–3–1
E-mail: †mine@gavo.t.u-tokyo.ac.jp

あらまし　言語音に対する新しい音響的表象手法を提案する．言語音を表現する場合，各音を生成する調音運動を参照する表現（調音音声学），及び，各音が持つ音響特性に着眼する表現（音響音声学）が広く用いられている．ここでは，対象とする言語音全体が成す系に着眼した物理モデリングを提案する．各音のもつ音響属性をケプストラムの分布として確率論的にモデル化し，任意の二音間距離をバタチャリヤ距離により情報論的に数値化する．こうして得られた音間距離行列は，言語音全体が成す系を「構造」として捉えることに相当するが，この構造は音声生成・収録・伝送・再生・聴取過程において不可避的に混入する乗算性及び線形変換性歪みに対して不変となる．乗算性歪みは構造のシフトとして，線形変換性歪みは構造の回転として観測される．即ち提案手法は，乗算性・線形変換性歪みを表現する次元を消失する形で定義される，言語音の音響的表象である．提案手法を語学学習者音声に適用した場合，声道長差異，個人性，マイク・録音室・伝送路特性，聴覚特性が消え，主に発音の母国語依存性のみが表現されることとなる．
キーワード　音声の普遍構造，音声学・音韻論，発音学習，CALL，学習者モデリング

Extraction of universal structure from speech and structural representation of language learners

Nobuaki MINEMATSU†

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

Abstract　A novel method of speech representation is proposed. Conventionally, speech sounds are represented by referring to their articulatory and acoustic properties (articulatory and acoustic phonetics). In this paper, an entire system of all the kinds of speech sounds of a language is physically characterized. After modeling the individual sounds of a language as distributions of cepstrum vectors, distance between any two of the sounds is quantified based on information theory. The resulting distance matrix of the sounds contains full information of structure of the system and the structure is found to be invariant with any kind of multiplicative and linear transformational distortions, which are inevitably involved in speech production, recording, transmission, and hearing process. Multiplicative distortion corresponds to shift of the structure and linear transformational distortion corresponds to rotation of the structure. The proposed method realizes speech representation with no dimensions to indicate the above two distortions. With the method, individual language learners are modeled, where properties of age, gender, size, mic, room, line, etc are unseen and only dependency of the pronunciation on their mother tongues is observed.
Key words　universal structure, phonetics & phonology, pronunciation training, CALL, learners' modeling

1. はじめに

周知のように，日本語と英語は音声学的（リズム，イントネーションを含む）にも言語学的にも非常に大きな差異を持つ言語対である [1]〜[4]。これに加え，各言語を母語とする話者が持つ音声知覚プロセス間における差異 [5], [6]，物事を考える発想，論理プロセス間における差異 [7]，更には，コミュニケーションにおける文化的差異 [8], [9] までもが存在し，日本人学習者にとって英語の習得・運用は非常に高いハードルとなっている。これらの状況を抜本的に改善する試みの一つとして，2002

年より文部科学省にて「英語が使える日本人」育成のためのプログラム [10] が開始しており，小・中・高・大の英語教育関係者による積極的な議論が展開されている．

近年の計算機技術・音声情報処理技術の進展に伴い，これらの問題解決に対する技術的サポートも盛んに議論されるようになった．2002 年度までの 3 年間に渡り，科学研究費補助金特定領域研究 (1)「高等教育改革に資するマルチメディアの高度利用に関する研究」の支援の下，多くの音声研究者，語学教育者により CALL（Computer Aided Language Learning）システムの研究開発が行なわれた [11]。また，日本人学生による大規模な読み上げ英語音声データベース（ERJ, English Read by Japanese）の構築も行われ，研究開発の効率化を促進した [12]．

音声情報処理を利用した語学学習支援を考える場合，対象は発音能力と聴取能力の向上となる．上記プロジェクトにおいて構築された発音評定システムの多くは，音声認識技術を用いて母語話者音響モデルとの照合を行ない，その結果に基づいた発音誤り検出や発音スコア算出をするものが多い．聴取能力向上を意図したシステムも，各種情報源に存在する英語音声（及びその書き起こし）に対して，音声とテキストとの同期をとって提示する形式のものが多い．このような学習支援は，アイデアそのものは古くから検討されてきたものであり（PC やビデオ教材として市販されている），最新の音声技術・計算機技術を用いてシステムの再構築を試みた例である．言い換えれば，高精度・高速化された種々の要素技術を用いて，問題の “量的解決” を試みた例として位置付けることができる．

日本人の英語運用能力の低さは，量的解決を意図した方法論で決着するのだろうか？日本人の類い稀なる勤勉性を考慮すると，量的解決で決着するのならば，既に問題は解決しているはずであると筆者は考えている.「英語が使える日本人」プログラムでは，英語教育の “質的変化” に関する議論が展開されている．技術支援に求められるべくは，その質的変化を支援する提案，更には技術によって初めて可能となる新しい教育論の提案も含まれると筆者は感じている．例えば，近年の語学教育では，native-sounding な発音ではなく，intelligible な発音が求められている [13]．科研プロジェクトによる発音評定システムでは，ほぼ全てにおいて，母語話者音響モデルとの音響照合に基づくスコアリングを行なっており，教育現場が求める “もの” とのずれは否めない．更に，音声認識技術は常に “sheep and goat”（利用者環境とシステム間の相性問題）を伴う技術である．もし不安定性が本質的に回避できない場合,「そのような技術を教育に導入してよいのか」と問いかけた場合，諸手を上げて歓迎する教育者・親がどのくらいいるだろうか?最近になって，CALL システムに懐疑的な報告を耳にするのも事実である [14]．

筆者は学生時代，英語劇を通して，口・腹周りの筋肉武装，腹式呼吸の習得から発音に取組んだ一人である（舞台では，母語話者よりも母語話者らしい発音を求められた）．また，そのような発音に向けた指導も試行錯誤的ながら行なった経験を持つ．その筆者が現在の学生教育の現状，英語教育者を目指す者に対する教育の現状，更に，音声情報処理技術の現状を考慮した上で，英語教育の質的変化に対する一つの提案を行なう．

2. 科学無き発音教育

2.1 音声学・音韻論は本当に発音教育に役立つのか？

CALL システムが呈する不安定性に関する報告が行われている [14]．この不安定性は，音声認識技術の利用者（即ち開発者，教師，学生）に起因する場合もあるが，音声認識技術そのものに起因する場合もある．教育（特に低年齢者に対する教育）の場で使われる技術は，まずその安定性が求められる．その意味において上記の報告例は「発音教育を真の意味で支援する技術はどこにあるのか？」という問いかけとして受け止めることができる．しかし，模索すべきは “技術” だろうか？筆者は「その前に，科学すら無いではないか」と感じている．

英語教師を目指す学生は，発音教育のための基礎知識として音声学・音韻論といった科学を習得する．音声学は言語を構成する個々の音を記述する科学であり，音韻論は一言語内の音のセットや並びに内在する関係，規則を記述する科学である．これらの科学は英語という言語の音やその規則を「紹介」するためには非常に優れたツールである.「紹介」することが「教育」であるならば，現在行われている英語教育は何ら問題を抱えていない．優れた「紹介」をしていると筆者は考えている.「紹介」だけでは意味がない，とするならば何が必要か？それは，学習者個人を記述する科学である．学習者が現在どういう状態にあるのか，その状態にある学習者にはどのような訓練が効果的であるのか，こういった教育を実現するためには，学習者を記述する「文字」が必要である．音声学や音韻論における議論の対象は，元来学習者個人を記述するための「文字」ではない．

音響音声学＋ PC というツールを使えば，学習者の発声を波形・スペクトル表示することが簡単に出来るようになった．これは一つの文字であるが，雑音の多い文字である．波形・スペクトルから得られる情報には，発音の善し悪し以外に，性別，年齢，体格，マイク特性，伝送路特性など種々の情報がある．これらは全て発音学習から見れば雑音であり，結局音響音声学が提供する文字は，学習者にとっては「雑音混じりの文字」となる．そして音声工学（音声認識）はこの雑音の多い文字を基盤とする技術である．不安定性が回避できないのも無理は無い．

2.2 発音教育が必要とする科学に対する考察

発音教育はどのような文字を必要とするのだろうか？本節ではこの文字に対して望まれる仕様について考察する．

学習者の「今」を記述可能. 上記した通りである．但し，その文字は一切雑音を含んではいけない．また，学習者（その親も含めて）が十分理解可能な文字でなければならない．

安定に動く技術が構築可能. 相性問題が原理的に生起し得ない技術を可能とする文字でなければならない．当然子供・高齢者を対象としても問題なく動作する技術を可能とする文字である．

現実的な学習目標を設定可能. 学習者が全て舞台役者を目指している訳ではない．通じる英語を目標とした時に，その学習目標を適切に記述できる文字である必要がある．最終的な学習目標は学習者自身が決定する必要があるが，学習者の学習意欲に応じた目標を適切に記述できる文字でなければならない．

調音音声学との連携が可能. 発音学習は最終的には，調音器官

をどう動かすか，といった議論に落ちる．その意味で，調音音声学と連携がとれる文字でなければならない．

以下，使用したデータベースについて述べ，その後，本論文，及び後続する論文 [15], [16] において，新しい文字（科学）を提案し，その科学の上に構築した技術を述べる．なお，発音の分節的側面のみに着眼した議論を行なっていることを断っておく．

3. 日本人による読み上げ英語音声データベース

ERJ（English Read by Japanese）データベースは，現在の音声情報処理技術水準，及び英語教育要項の両側面からの要請を考慮して作成され，文・単語セットの選定，話者の選定，収録方式において種々の工夫が行なわれている。同一読み上げセットの米語母語話者音声の収録，一部の日本人英語音声データ（約 3,800 文音声，約 5,700 単語音声）に対する複数の米語母語話者英語教師による評定ラベリングも行なわれている [12]。本データベースの特徴は，まず，収録時に種々の発音情報を参照した収録を行なっている点である．音素（発音）記号，リズムやイントネーションを表す記号の他に，一部の文には文意を明記した上で読み上げリストを作成している．そして発声者は事前の発音練習が許されており，収録時も「自らが正しいと考える」発音ができるまで繰り返して発声させている．即ち,「日本人学生が正しいと考える」英語音声のデータベースとなっている．発声者の選択も日本各地の大学・高専から準ランダムに抽出された話者を使用しており，非常に幅広い英語発音習熟度の話者の音声資料が収録されている．話者数は男性 100 名，女性 102 名である．本研究（[15], [16] を含む）の目標は，これら 202 名の「今」を記述する発音カルテの作成と，彼らに与える 202 種類の異なる薬の処方に相当する．語学教師はこのカルテの記述法と薬の処方法を，実際に教壇に立ってから学生との接触を通し，時間をかけて自らに蓄積していく．本研究は彼らが発音教育に対して蓄積する経験と知恵を，知識データベースとして蓄積する術を提供する枠組みとして捉えることもできる．

4. 発音教育のための音声モデリング

4.1 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にした工学である．音響音声学は音声学の一分野であり，音声学は言語の種類を問わず，言語音の一つ一つを何らかの観点から記述することを目的とした科学である．言語音の調音的性質に着眼した調音音声学と音響的性質に着眼した音響音声学がある(注1)．音響音声学における音の記述を工学的・数理的にモデル化することで，音声認識で広く使われる音響モデルが実現される．

CALL システムのみならず，音声認識技術に基づくシステムの不安定性は常にユーザの苦情にさらされるのが現状である．なぜユーザはシステムが不安定だと感じるのか？不安定要因の原因を音響モデリングに模索した場合，mismatch 問題がその対象となる．音響モデルの学習環境と実際の使用環境が異なれば，何らかの不整合が生じ，予期せぬ結果を生むこととなる．

(注1)：これら以外に聴覚音声学という分野があるが，ここでは言及しない．

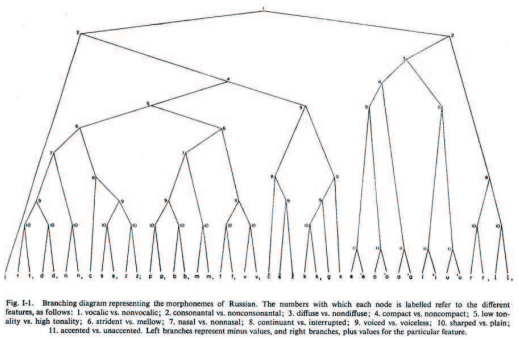

Fig. I-1. Branching diagram representing the morphonemes of Russian. The numbers with which each node is labelled refer to the different features, as follows: 1. vocalic vs. nonvocalic; 2. consonantal vs. nonconsonantal; 3. diffuse vs. nondiffuse; 4. compact vs. noncompact; 5. low tonality vs. high tonality;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

この不安定要素を回避するために，例えば不特定話者・環境モデルが構築されたり，また，少量のデータを使った話者・環境への適応技術，更には話者・環境性を消す正規化技術が検討されている．しかし，これらの技術を駆使したとしても，最終的に得られるモデルや音響特徴量には，話者性や環境特性が（例えば）“1” となるだけであり，それは，“1” という特性を持った話者や環境であることに他ならない．つまり，話者・環境特性を表現する次元は存在したままである．**第 2.2 節**で述べた「雑音を含まない」音声表現を完全に実現するためには，雑音を “1” にする技術を求めるのではなく，その雑音を表現する次元を保有しない音声表現によって完成されると考えるべきである．

そのような都合のよい表現方法が存在するのだろうか？話者が異なれば声道長は異なる．収録場所が異なれば，マイクの特性も異なる．バーク尺度で近似される聴覚特性も個人間差異は存在する．即ち，生成・収録・伝送・再生・聴取という行為そのものが不可避的に歪みを混入させる．加算性雑音と異なりこれらの歪みは「歪んでいない音声が存在しない」という意味で原点が存在しない（加算性雑音であれば，その雑音源を抹消すれば雑音はゼロ（即ち原点）となる）．このような不可避的な歪みが混入している音声に対して，個々の「音」を記述しようとすれば，そこには何らかの値の歪みが入る．つまり，本研究で求める音声表現は音声学を基盤にした考えでは実現不可能な表現である.「音声学を捨てる」これが本研究の第一歩である．

4.2 個々の音が織りなす系を記述する方法論

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．音韻論は言語音を抽象化して “音素” として捉える．即ち，異音の存在を許容しない．当然，話者・収録環境の違いなども許容しない(注2)．音韻論は，ある言語を構成する音素の集合に内在する規則・関係，或は，その音素の並びに内在する規則・関係を議論する．音素（やその関係）が言語音の物理現象を抽象化して得られる概念であるとするならば（音声学による音記述を抽象化して音韻論による音記述が生まれるのであれば），**第 4.1 節**で述べた「歪みを表現する次元を保有しない」音声表現は，音響モデルに対して抽象化（音韻論化）をかけることが，その実現に向けた一つのアプローチであると考えられる．

(注2)：厳密に言えば，IPA 記号がそうであるように，音声学でも話者・収録環境の違いは考慮しない．つまり音声学も音声の物理をある程度抽象化して捉える科学として捉えることができる [17].

音韻論における系の記述の一例として，Halle によるロシア語の音素樹型図を**図 1** に掲げる [18]．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ぶか，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持つように弁別素性が選ばれる（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，素性と言語現象に着眼した音素分類は Jakobson によって提唱されたが [19]，その根底には Saussure の構造主義の哲学がある [20]．

Blache はこの音素群が成す構造に着眼して，言語獲得過程を説明している [20]．つまり，音を獲得するのではなく，音の中に潜む構造を獲得するのが言語獲得の過程であると考える．このように音の構造への着眼は，言語獲得の議論の中で使われたツールであり，外国語学習においてもその有用性は高いと言える．更に，**図 1** に示した音素樹型図が，人が自らの母国語に対して無意識的に持ってしまっている言語音感の具体的表象であるとするならば，この表象は，母語話者間では普遍的な構造であり，この構造を自らの口に宿すことが，外国語学習における発音学習の最終目標として位置づけることができる．そして，学習者個人に対する音韻論的記述が可能であれば，学習者の記述と母語話者の記述を比較することで，最終目標（母語話者が持つ言語音感）との構造的歪みを議論することが可能となる．

Halle が行なった作業を学習者個人に対して行なうことは不可能である．素性と音素のマッピングは学習者によって異なり，学習者個人別に言語現象を求めることはできない．Halle の行なったクラスタリングは，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，それに対して，言語・音声知識を用いて，適切な二分割を繰り返す，というものである．このトップダウンクラスタリングを行なうことで，自然類のような，言語現象との対応をとることができる．音素クラスタリングにおいて，言語現象との具体的な対応をとることを諦めた場合，即ち MLLR で行なわれるようなボトムアップクラスタリングを考えた場合，学習者個人に対して音素樹型図が求まることとなる．この場合必要となるのは言語・音声知識ではなく，任意のリーフノード間の距離だけが必要となる．以下，音響モデル間の距離だけに着眼して学習者別に構築される音素樹型図を考察するが，その中で，音声の物理に内在する美しい普遍構造が呈することとなる．

4.3 静的歪みを表現する次元を持たない音声の音響的表現

4.3.1 話者依存 HMM の学習

ERJ データベース中の全話者（日本人 202 名，米国人 20 名）に対して，話者別に HMM を学習した．音素バランス文セットの中のサブセット（60 文）を各話者が読み上げており，この 60 文を用いて monophone を学習した．なお，読み上げたサブセットは話者によって異なっている．60 文という少量データでの学習のため，話者によっては一部二重母音が含まれない場合があり，二重母音は学習対象から除いた．その結果 34 音素を，各話者毎に学習した．**表 1** に学習条件について示す．なお，学習時に使用する音素書き起こしであるが，PRONLEX を参照して作成し，発音誤りは一切反映されていない．

表 1　HMM 学習条件
Table 1　Acoustic conditions for training HMMs

サンプリング	16bit / 16kHz
窓	窓長 25 msc，シフト長 10 ms
パラメータ	メルケプストラム（0〜24 次元）及びその動的特徴
話者	日本人 202 名，米国人 20 名
学習データ	一話者当り 60 文（音素バランス文の一部）
HMM	環境非依存の 1 混合 monophone（対角分散行列）
トポロジー	5 状態 3 分布
音素	b,d,g,p,t,k,jh,ch,s,sh,z,zh,f,th,v,dh,m,n,ng,l,r,w,y,h, iy,ih,eh,ae,aa,ah,ao,uh,uw,er,ax

4.3.2 音素モデルのボトムアップクラスタリング

上記したように，ボトムアップクラスタリングは，対象とする要素群に対して，任意の要素間の距離が求まれば行なうことができる．以下，音素モデルをケプストラムベクトルによって構成されるガウス分布であると仮定して（即ち GMM で音素をモデリングすることに相当する）話を進める．本研究では音素間距離としてバタチャリヤ距離の平方根(注3)を用いる．分布 u と v 間のバタチャリヤ距離は以下の式によって与えられる．

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

μ_u は u の平均ベクトル，μ_{uv} は $\mu_u - \mu_v$ を，Σ_u は u 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，$P_u(\boldsymbol{x})$ と $P_v(\boldsymbol{x})$ 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．言い換えれば,「二つの事象 u と v 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**式 (1)** である．

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（即ち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．そして樹型図の生成は，構造を視覚化する手段の一つである．ここでは，Ward 法を用いたクラスタリングを考える．Ward 法は，統合により生じる累積歪みが最小となる二要素を順次統合していく方法である．最終的に得られる樹型図の高さは，要素群を一点で代表させた時の歪み，即ち，VQ 歪みに相当する量になる．このようにして，202 人の学生の「今」が「木」という形で表現されることになるが，各学習者の木の様子を具体的に見る前に，まず，本分析手法の特徴を考察する．

(注3)：平方根を使う理由は文献 [15] を参照して戴きたい．

4.3.3 構造として捉えられた音声事象が持つ特徴

さて，音素間距離行列は，各音素の相対的な関係を「距離」というスカラー量にまで落として記述する方法論であり，音声の音響情報の多くが損失しているのは明らかである．以下，何が失われているのかについての考察を行なう．まず，各音素モデルを学習する時に使用した学習データ（ケプストラムベクトル c）に対して，共通の一次変換をかけることを考える．

$$c'_t = Ac_t + b, \tag{2}$$

この変換により，平均ベクトル μ と分散共分散行列 Σ は以下のように変換される．

$$\mu' = E(c'_t) = A\mu + b \tag{3}$$

$$\Sigma' = E(c'_t - \mu')(c'_t - \mu')^T = A\Sigma A^T \tag{4}$$

二つのカテゴリ（分布）u，v を考える．上記の変換式は，各分布を構成するデータに対して如何なる一次変換を施しても，バタチャリヤ距離がその前後で不変であるという性質を導く．

$$\begin{aligned} &BD(\mu'_u, \Sigma'_u, \mu'_v, \Sigma'_v) \\ &= BD(A\mu_u + b, A\Sigma_u A^T, A\mu_v + b, A\Sigma_v A^T) \\ &= BD(\mu_u, \Sigma_u, \mu_v, \Sigma_v) \end{aligned} \tag{5}$$

これは，**第 4.3.2 節**で考えた音素間距離行列は一次変換に対して不変であることを意味し，最終的に，音声に対して確率論・情報論的に定義される構造は不変であるという性質を導く．

広く知られている様に，一次変換 $Ax+b$ はアフィン変換と呼ばれ，画像処理の世界では，ユークリッド空間における構造（即ち図形）に対して種々の変形を施す際に用いられる．例えば，行列 A を掛けることによる変形は，構造に対する，軸毎に独立なスケーリング，せん断（ずらし），回転といった要素に分類される．一方ベクトル b を足すことによる変形は，構造のシフトとして観測される．これらの変形要素を組み合わせることで多種多様な構造変形が実現される．さて，ユークリッド空間では構造は点の集合として扱われるが，この点が平均ベクトルであり，その平均ベクトルには分布が付随している状況を考える．この場合，二点間（即ち二分布間）の距離はバタチャリヤ距離で測定する．分布を構成するデータに対してアフィン変換をかけた場合，構造はどのように変化するだろうか？上記した変形要素のうち，任意の二点間の距離を変えない変形は回転とシフトのみである．つまり，行列 A を掛けることによる変形は，各点に付随する分布を考慮し情報理論に基づいて距離を算定した場合，構造の回転として観測されることとなり，b を足す変形は，シフトとして観測されることになる．

ケプストラム係数の一次変換は，音響音声学的にどのように解釈されるのだろうか？b を足す演算は，b で表現される対数スペクトルを足す操作であり，これはフィルタリングに相当する．マイク，伝送路の特性，更には録音室の音響特性の一部もこれに相当する．GMM による話者モデリングが，該当話者の長時間スペクトル系列の平均値をモデル化していることが考えれば，話者性の一部もフィルタリングとして解釈することが可能である．これらは全て構造のシフトとして観測される．さて，A は何を表すのか？声道長の話者間差異は，スペクトルのフォルマントシフトとして観測され，音声認識の世界では，スペクトルに対して周波数ウォーピングをかけることで，このシフトを近似することが広く行なわれている．文献 [21] は，連続かつ単調であれば，任意の周波数ウォーピングはケプストラム領域では全て A を掛ける演算に変形されることを数学的に導いている．よってこのウォーピングによる音声の変形は，構造の回転として捉えられることになる．声道長の伸び縮みを人間の成長と考えれば，人間の成長によってその人間の発声する音声の構造は回転していることになる．周波数ウォーピングは A となるが，その逆は真ではない．即ち，異なる音声変形が A となる可能性は多分に残されている．何れにせよ，A として実現される音声変形は構造の回転であり，構造そのものは不変である．ガラテアプロジェクト [22] を通して配付されている HMM 音声合成器 G-talk では，周波数ウォーピングを通して同一話者の HMM から異なる声質の合成音を出すことが可能であるが，あの声質変化は構造の回転となる．以下，音声の物理に対して定義されるこの構造を「発声者及び聴取者の生理学的条件，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，音声の音響的普遍構造と定義する．

普遍構造の存在を人間間の音声コミュニケーションにおいて考察する．話し手の音声が何らかのメディアを通して聞き手に聴取される過程において，**第 4.1 節**で述べたように，生成，収録，伝送，再生，聴取のいずれの過程においても不可避的に歪みが混入する．しかし，音声の中に構造として存在可能な情報は（完全なる消失が可能な加算性雑音を無視すると），この過程において一切歪みを受けることなく聞き手（の脳に）に伝達されることになる．つまり，話し手と聞き手の間に，情報の損失・改変が，原理上起こり得ないコミュニケーションチャネルが存在することになる．「このチャネルによってどのような情報が伝達可能なのか」という議論は今後の研究課題となるが，従来行なわれてきた音声科学・工学の議論の中で，このような完全なチャネルが存在することを示した例を筆者は知らない．

音声認識の世界では広く不特定話者音響モデルが使われる．異なる話者の音声から構築した音響モデルである．当然，個々の話者の音声には普遍構造が（異なる位置，異なる向きで）存在しているが，それを混ぜ合わせれば，当然構造はぼけてくる．上記したチャネルの存在が既知であった場合，不特定話者音響モデルというものを工学者は構築しただろうか？幸いなことに近年になって，不特定話者モデルにおける「不特定話者性」を解消する音響モデル構築が行なわれるようになった．学習時に一次変換に基づく話者適応をかけ，架空の話者性を想定し，その話者の特定話者モデルを作り，認識時には，再度一次変換にて入力話者の音声に適応させる枠組みである（Speaker Adaptive Training, SAT [23]）．この SAT を普遍構造の観点から眺めると，複数の学習話者が同一文セットを同一の発話スタイルで読み上げたとるすと，SAT は無意味である．各話者の音声データ（構造）は一次変換（回転とシフト）によって全く同一のものとなるからである．SAT は複数の話者による，異なる文セッ

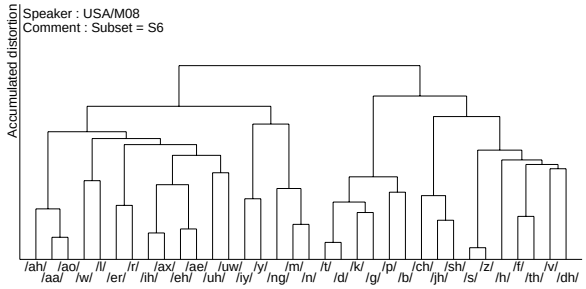

図 2　米語母語話者による音素樹型図

Fig. 2　An English phoneme diagram made from a native speaker of American English

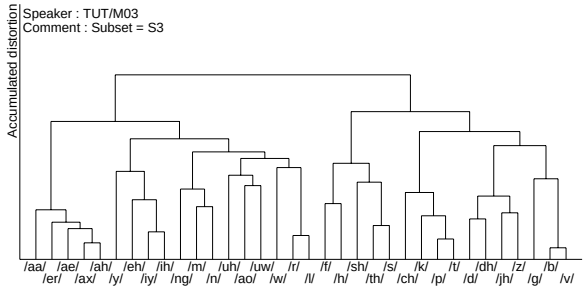

図 3　日本人による英語音素樹型図

Fig. 3　An English phoneme diagram made from a native speaker of Japanese

ト，異なる発話スタイルをあたかも同一話者が読み上げたように見なしながら音響モデルを学習する方法である．

MLLR における話者適応では，不特定話者モデルに対して，混合レベルのボトムアップクラスタリングをかけ，混合を幾つかに分類した上で，各クラスに対して別個の一次変換をかける．上記した普遍構造の議論は変換行列数 1 の MLLR に相当する．不特定話者モデルの場合，モデルセットの各部分は異なる個人性で構成されていると考えられ，それらを特定話者に適応する場合，当然複数の変換行列が必要となる．これは複数の話者による音声を元に HMM 合成用のモデルを構築する場合，この個人性の違いを解消する必要が生じる [24] ことからも理解できる．

5. 普遍的構造抽出に基づく学習者モデリング

前節で，音声生成・収録・伝送・再生・聴取過程において不可避的に混入する静的歪みを表現する次元を持たない音声の物理表象が可能であることを示した．本節では紙面の都合上，提案手法に基づく学習者と母語話者の構造の例を一名ずつ示すに留める．なお，音素間距離は 3 つの状態間距離の平均値で定義している．他の学習者の音声とその構造との対応に興味がある読者は，直接筆者と連絡をとって戴きたい．**図 2** に母語話者，**図 3** に日本人学生による構造の例を示す．前者が音声学的に非常に “健康的な” 構造を呈しているのに対し，後者が広く知られる日本人英語の癖を見事に表現している様子が分かる．

6. ま と め

本研究では，音声に不可避的に混入する静的な歪みを表現する次元を消失させた音声の音響的表象が可能であることを示し，それに基づいて語学学習者を記述することを提案した．本手法は言語現象との関係を一切断っている点で音韻論とは異なり，また単音の絶対的音響特性を何ら表現し得ないという点で音声学とも異なる，全く新しい音声の音響的表象手法である．本手法を用いることで，語学学習者の「今」を全く歪み無く表現することが可能となるが，提案手法に基づく語学学習応用の実際については，文献 [15], [16] を参照して戴きたい．

文　献

[1] 川越，英語の音声を科学する，大修館書店，東京 (1996)

[2] 小川，理屈で分かる英語の発音，ノヴァ・エンタープライズ，東京 (1999)

[3] 深澤，英語の発音パーフェクト辞典，アルク，東京 (2000)

[4] F. Ramus, “Acoustic correlates of linguistic rhythm: Perspectives,” Proc. Speech Prosody, pp.115–11 (2002)

[5] K. Tajima, *et al.,* “Perceptual learning of English syllable rhythm by young and elderly Japanese listeners,” Proc. General Meeting of the Phonetic Society of Japan, pp. 103–108 (2002)

[6] A. Cutler, “Listening to a second language through the ears of a first,” Interpreting, vol.5, no.1, pp.1–23 (2001)

[7] 松本，英語で考える本，英友社，東京 (1968)

[8] R. B. Kaplan, “Cultural thought patterns in inter-cultural education,” Language Learning, vol.16, pp.1–20 (1996)

[9] 淺間，“英語教育における「異文化コミュニケーション」の普遍的視点をめぐって”，静岡大学情報学研究，vol.3，pp.1–10 (1997)

[10] http://www.mext.go.jp/b_menu/shingi/chousa/shotou/020/sesaku/020702.htm

[11] 文部科学省科学研究費補助金特定領域研究 (1) 「高等教育改革に資するマルチメディアの高度利用に関する研究」平成 14 年度研究成果報告書 (2003)

[12] 峯松他，“英語 CALL 構築を目的とした日本人及び米国人による読み上げ英語音声データベースの構築”，日本教育工学会論文誌 (2004，採録決定)

[13] D. Crystal, English as a global language, Cambridge University Press, New York (1995)

[14] A. Neri, *et al.,* “Automatic speech recognition for second language learning: How and why it actually works,” Proc. ICPhS, pp.1157–1160 (2003)

[15] 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資料，SP2003-180 (2004)

[16] 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)

[17] 前川，“音声分野における統計的研究手法 —音声の変異をめぐって—”，日本語学，vol.20，pp.144–156 (2001)

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

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

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

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

[22] 嵯峨山他，“擬人化音声対話エージェント基本ソフトウェアの開発プロジェクト報告”，電子情報通信学会音声研究会資料，SP2003-168, pp.73–78 (2003)

[23] T. Anastasakos, *et al.*, “A compact model for speaker-adaptive training,” Proc. ICSLP’96, vol.2, pp.1137–1140 (1996)

[24] J. Yamagishi, *et al.*, “A context clustering technique for average voice models,” IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.534–542 (2003)