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

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　音声を用いた人間間のコミュニケーションでは，その生成，収録，伝送，再生，聴取の何れの過程においても，乗算性，或は線形変換性の歪みが不可避的に混入する．しかし，音声（音素群）を情報理論的に構造として捉えた場合，その構造はこれら歪みに一切影響を受けず伝搬されることが峯松により示されている．即ち，情報の損失，改変が原理的に生起し得ないコミュニケーションチャネルの存在が明らかとなっている．提案されている音声表象は，乗算性・線形変換性の歪みによって不変であるため，この表象のみに基づく音声アプリケーションは，その安定性が飛躍的に向上することは自明である．本研究ではこの表象を語学学習支援に応用する．母語話者の発音を構造として捉えた場合，それは話者，収録環境に依存しない，nativeness そのものの音響的表象となる．そこで，母語話者より抽出される構造と学習者から抽出される構造とを比較することで発音の自動評定を試みる．と同時に，抽出される表象には発音習熟度のみが残り，話者性・収録環境差異が完全にそぎ落とされていることを実験的に示す．
キーワード　音声の普遍構造，音声学・音韻論，発音学習，CALL，自動評定

Automatic scoring of language learners' pronunciations based on the distortion of their universal structures

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　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. Minematsu showed that, if speech is represented as structure based on information theory, these distortions cannot change the structure. In other words, he proved that there is a perfect communication channel between a speaker and a hearer on which any piece of information cannot be modified or lost. If speech application is possible only with the proposed representation, robustness of the application must be remarkably improved. In this paper, this representation is applied to CALL (Computer Aided Language Learning). Since the structural representation of speech is thought to include no speaker individuality or recording conditions, the representation of all the phonemes obtained from a native speaker can be viewed as the nativeness. Automatic scoring of the pronunciation is investigated by comparing native structure and a learner's structure. Also, it is experimentally shown that the proposed representation include completely no information of speakers' identity and recording conditions.
Key words　universal structure, phonetics & phonology, pronunciation training, CALL, automatic scoring

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），言語学的，更には認知科学的にも非常に大きな差異を持つ言語対であり [1]〜[6]，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 [7] なども開始されており，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．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ており [8]，また，効率的な研究開発を促進するために，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，近年行なわれた [9].

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．本研究では，これらのうち発音能力の向上を支援する環境について検討する．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形態となっている．しかし周知のように，現在の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．日本の場合，初等教育に英語教育がシフトしようとしているが，一方で英語教育者の不足が懸念されている．そのような場に技術が代替案の一つとして入ることが予想されるが，子供音声の処理は現在の音声認識技術が最も苦手とする対象の一つである．もし不安定性が本質的に回避できないとすると,「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？大学生を相手とした場合，学習者の思慮分別を期待したシステム作りができるだろうが，初等教育において，不安定性を解消できていない技術を導入することは，技術者として大きな不安を感じざるを得ない．実際，最近になって，CALL システムに懐疑的な報告を耳にするようになった [10].

音声認識技術の不安定さ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．この mismatch 問題の完全な解決は，話者性，収録特性，伝送特性といった発音習熟度とは無関係の音響現象を表現する「次元」を保有しない音声の音響的表現方式を実現することで達成される．このような音声の表現方式が峯松によって見い出された [11]．音声現象を確率論的に有限個の状態として記述し，任意の状態間距離を情報論的に算出し，最終的に音声を状態セットが成す構造として捉えると，その構造は，如何なる乗算性，線形変換性歪みに対しても不変となる．この構造を，音声に内在する音響的普遍構造と定義する．そこでは，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，音声生成，収録，伝送，再生，聴取の何れの過程においても不可避的に混入する静的な歪みが原理的に影響を及ぼすことができない，完全なコミュニケーションチャネルが存在することが示されたことになる [11].

母語話者の音声からその発音を構造として表現すると，それは話者性，収録特性が消え去り，残るものは nativeness そのものとなる．本研究では，母語話者より抽出される構造と学習者から抽出される構造とを比較することで，発音の自動評定が可能であることを示す．と同時に，構造抽出という過程が完全に話者性，収録特性をそぎ落としていることを実験的に示す．

2. 音声に内在する音響的普遍構造

2.1 音声の音響的特徴の構造化

ある言語に含まれる各音素を，ケプストラムベクトルによって構成される多次元ガウス分布であると仮定する．さて，音素間距離をバタチャリヤ距離の平方根(注1)で定義する．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．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．即ち，任意の音素分布間距離を上式で求めた場合，それは，音素群が成す構造を情報理論的に規定したことに他ならない．

2.2 構造として捉えられた音声事象が持つ特徴

各音素分布を構成する各データ（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た場合の構造変化を考察する．

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

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．

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

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

さて，二つのカテゴリ（分布）u，v を考える．上記の変換式は，分布間のバタチャリヤ距離が一次変換前後で不変であるという性質を導く．

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

即ち，バタチャリヤ距離は，各分布を構成するデータに対して共通の如何なる一次変換を施しても分布間距離は不変となる．これは，音素分布間距離は如何なる一次変換に対しても不変であることを意味し，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

(注1)：平方根を使う理由は**第 3.2 節**で述べる．

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，各点が有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，行列 A を掛けることによる変形は構造の回転として観測されることとなり，b を足す変形は構造のシフトとして観測されることになる．

3. 構造の比較に基づく外国語発音の自動評定

3.1 音素間の相対的距離関係のみに基づく音声表象

筆者が提案する音声表象は，音素間の相対的な関係のみを，距離というスカラー量にまで落として行なわれる．その結果，各音の絶対的な音響的特性は一切見えなくなる．つまり，単音の認識も単音の合成も一切不可能である．音声アプリケーションが音声認識・音声合成の要素技術に基づかねばならないとしたら，本音声表象のみでは何もできない．しかし，音のモデルに基づく従来の音声情報処理に一端戻ってしまうと，「不安定性」が不可避的に付きまとう．本論文と後続する論文 [12] では，筆者が提案する「関係のみによって定義される音声表象」で，発音の評定のみならず，矯正すべき発音部位に対して，その優先順位が導出されることを示す．これは，従来の音声情報処理の常識を覆す試みである．**図 1** に日本人が発声した英語 60 文から構成した音素樹型図の一例を示す [11]．以下，これと母語話者樹型図間の構造的差異を定量化する方法について検討する．

3.2 音素群より構成される構造間の差異の導出

まず，一つの（一人の）構造内に観測される特徴について考える．今，N 次元ユークリッド空間内に M 個の点があるとする（$\{P_i\}, 1 \leqq i \leqq M$）．任意の二点を結ぶことで，${}_MC_2$ の対角線が定義され，構造を考えることができる．この構造の重心を P_G とした場合，次式が真となる．

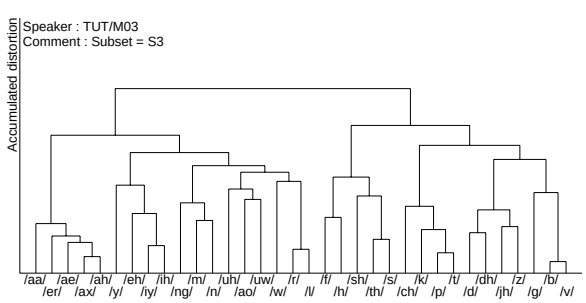

図 1　日本人による英語音素樹型図

Fig. 1　An English phoneme diagram made from a Japanese student

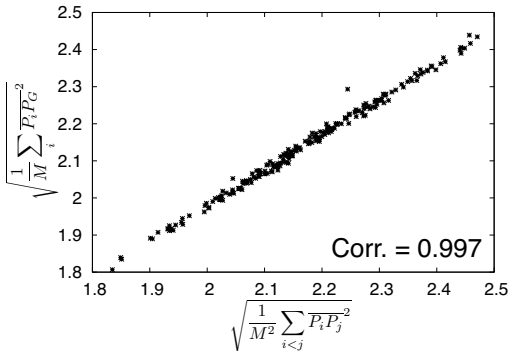

図 2　バタチャリヤ距離の平方根が持つ特性

Fig. 2　Properties of square root of Bhattacharyya distance

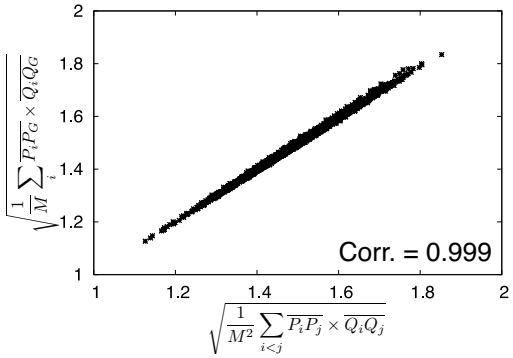

図 3　二つ構造間においてバタチャリヤ距離の平方根が持つ特性

Fig. 3　Properties of square root of Bhattacharyya distance found in two structures

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}^2} = \sqrt{\frac{1}{M}\sum_i \overline{P_iP_G}^2} \tag{6}$$

さて，提案する音声表象では二つの分布間距離は両者の同時確率密度から算出される自己情報量（バタチャリヤ距離）で定義される．上記のユークリッド空間における性質はバタチャリヤ距離では近似的にも成立しない．しかしバタチャリヤ距離の平方根を分布間距離尺度として使用すると，例えば，日本人 202 人が呈する英語単母音（11 種類）構造において，**式 (6)** の両辺は**図 2** に示す関係を呈する．全音素を用いても同様である．即ち，バタチャリヤ距離の平方根を使うと，**式 (6)** で示されるユークリッド性は近似的に満たされる．本研究でこの距離尺度を使用しているのはこれが理由である．

次に二つ（二人）の構造 $\{P_i\}$，$\{Q_i\}$ を考える．ただし，P_{i_0} と Q_{i_0} は同一音素である（構造を構成する点は P，Q 間で対応がとれる）とする．様々な構造歪みが予測される日本人英語の母音，全音素構造に着眼すると，以下の式が近似的に成立する．

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}\times\overline{Q_iQ_j}} \approx \sqrt{\frac{1}{M}\sum_i \overline{P_iP_G}\times\overline{Q_iQ_G}} \tag{7}$$

実際に 202 人の日本人に対して，同一文サブセットを読み上げた任意の 2 人に対して上式を求めたものが**図 3**（但し母音）である．全音素を使用した場合も同様の関係が得られる．

式 (6)，**式 (7)** より，以下の関係式が導かれる．

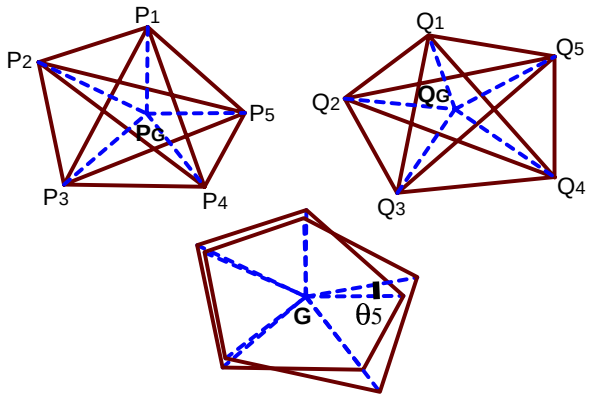

図 4　構造の比較に基づく音響的照合

Fig. 4　Acoustic matching based on structural matching

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \tag{8}$$

$$= \sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}^2 - 2\overline{P_iP_j}\times\overline{Q_iQ_j}+\overline{Q_iQ_j}^2}$$

$$\approx \sqrt{\frac{1}{M}\sum_{i}\overline{P_iP_G}^2 - 2\overline{P_iP_G}\times\overline{Q_iQ_G}+\overline{Q_iQ_G}^2}$$

$$= \sqrt{\frac{1}{M}\sum_{i}(\overline{P_iP_G}-\overline{Q_iQ_G})^2} \tag{9}$$

式 (8) は音素間距離行列をベクトルと見なし，行列間距離をベクトル間のユークリッド距離として求めたものに相当する．**式 (9)** の物理的解釈を考える．**式 (9)** は二つの構造の絶対座標系の中での位置，方向とは無関係に算出される量である．今，二つの構造の重心が重なるように両者をシフトさせる（その時の重心を G とする．**図 4** 参照）．その後，対応する点と点が重なるよう，いずれかの構造を重心周りに回転させ，$\sum|\theta_i|$ が最小となる方向を決定する．なお，θ_i は**図 4** にあるように $\angle P_iGQ_i$ である．ここで $|\theta_i|$ が十分に小さい場合は，

$$\left|\overline{P_iG}-\overline{Q_iG}\right| \approx \overline{P_iQ_i} \quad \text{（但し，シフト\&回転後）} \tag{10}$$

であり，回転&シフトによって $\sum_i|\theta_i|$ が十分に小さくなれば

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \approx \sqrt{\frac{1}{M}\sum_{i}\overline{P_iQ_i}^2} \tag{11}$$

が成立する．構造のシフトが乗算性歪みを，構造の回転が線形変換性歪みを表現することを考えると，**式 (11)** は上記両歪みに関する完全な適応をかけた後の音素間距離の平均値（の近似値）が，MLLR [13] や SAT [14] が行なう最尤推定を行なうことなく，構造と構造との比較によって導かれることを意味する．個々の音のモデルを保有せず，音素分布間の関係を距離というスカラー量にまで落とし，音素分布が成す構造だけの情報を用いた発音評定が可能である「からくり」が**式 (11)** である．

3.3　日本人・米国人に対する 2 話者間の構造歪み

式 (11) で定義された音素間距離行列のみを用いた 2 話者間の構造歪みと，単純なスペクトル距離に基づく 2 話者間距離を比較する．ERJ データベース [9] の各話者（全 222 名）に対して**表 1** に従って HMM を作成し，音素間距離を対応する状態間距離の平均値として定義し，音素間距離行列を算出した．なお，話者によって読み上げた文は異なる（8 つの文サブセットのいずれか）．音素構造は読み上げた文サブセットに依存する傾向があるため，以降の構造間比較は，いずれも，同一文サブセットを読み上げた二話者間で行なっている．

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

図 5 上図に，母音構造を用いた結果を示す．横軸は，二話者間の比較を構造歪みで算出したものである，縦軸は二話者間の距離を得られた HMM 間の距離の平均として求めたものである（つまり，音響空間内の位置に基づく距離）．構造歪みでは，日本人・米国人間（AE–JE）距離と米国人・米国人間（AE–AE）距離とが明確に分れているが，位置に基づく距離では殆ど重なっている．これは HMM が話者依存であるからだが，話者性の差異が大きな mismatch を引き起こしていることが分かる．一方，**図 5** 下図に全音素構造（但し二重母音を除く）を用いた結果を示す．この場合，構造歪みにおいても AE–JE 間距離と AE–AE 間距離が近づいてきている．これは 1) 本来音素 HMM

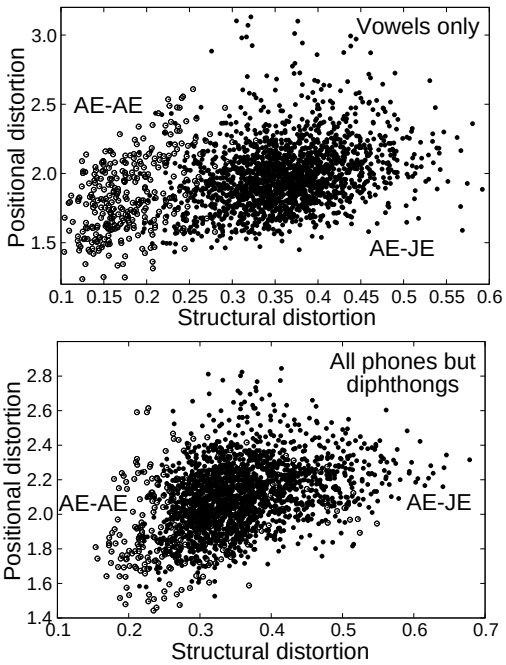

図 5　構造に基づく歪みと位置に基づく歪み

Fig. 5　Structural and positional distortions

は，その中に支配的な状態があるにも拘らず，音素間距離を，対応する状態間距離の単純平均で求めている，2) 構造の回転を実現するためには，分散共分散行列は全角である必要があるが，ここでは対角化分散行列を使用している，などの実験条件の不具合も否めないが，ここでは，**式 (11)** における構造歪み計算における音素対を適切に選択することで解消を図る．

3.4 学習者発音に見られる構造歪みに着眼した自動評定

どの音素対を選択するのかは，どのような応用を考えているのかに依存する．例えば，発音自動評定を高精度に行なうことを意図した場合以下の方法論が考えられる．ERJ データベースには分節的特徴に着眼した評定スコアが付けられている（5 点が最高，5 人の米国人教師によるスコアの平均値）．即ち，ある母語話者（教師）の構造に対して「教師・学生間の構造歪み」と「$5 - s_i$ （s_i は学生 i の評定スコア）」とが，より高い相関を持つように音素対を選択すればよい．ERJ データベースで全文サブセットを読み上げた男性話者を教師として選び，上記相関値が最大となるように音素対を逐一選んでいくと，52 音素対で相関は最大（0.88）となった．なお，教師と学生間で同一の文セットに対する HMM を用いて，構造歪みを計算している．**図 6** にその時の発音評定値と構造歪みとの相関の様子を示す．比較的良好な結果が得られていることが分かる．

3.5 日本人英語の分類

構造間の差異が定量的に求まれば，学生間の差異が定量的に求まり，学生間距離行列が計算できることとなる．それを使えば，202 人の日本人英語を分類し，日本人英語の「型」を定義することすら可能となる．そのための音素対選択としては以下の方法論が考えられる．任意の二人の学生 i, j の評定スコア差 $|s_i - s_j|$ と両者の構造歪みの相関を最大とする音素対を選択する．なおこの場合に限り，異なる文サブセットを読み上げた学生間においても構造歪みを計算した．前節と同様に相関値最大となるように逐一音素対を選ぶと，24 音素対で相関は最大となった（0.60）．この時の音素対を用いて 202 名の学生を分類したところ，教師による評定点が高い学生が樹型図内に固まって存在する様子が観測されたが，それと同時に，文サブセットによるクラスタリングの様子も同時に観測された．**第 3.3 節**でも述べたが，音素群が成す構造は，習熟度だけではなく，読み上げた文サブセットも影響を及ぼす．そこで，比較的幅広い

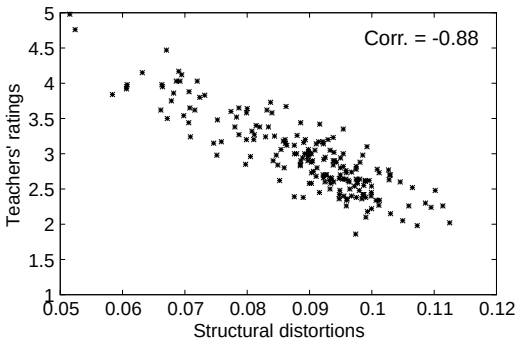

図 6　構造歪みに基づく自動発音評定結果
Fig. 6　Proficiency rating based on structural distortion

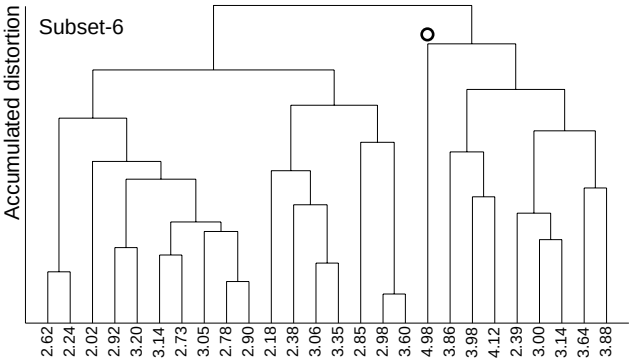

図 7　構造歪みに基づく日本人英語の分類
Fig. 7　Classification of Japanese English based on its structural distortion

発音習熟度の話者が集まった文サブセット 6 の話者のみを使って得られた学生樹型図を**図 7** に示す．英語教師による評定スコア平均も同時に示している．26 名の話者を 3 つに分類した時に，一人で 1 クラスタを構成している話者がいる．この話者は 202 人の中にいた唯一のバイリンガル話者である．それ以外の 25 名の様子を見ると，まず，発音能力の高いグループと低いグループとに分れ，後者が更に二分割されている様子が分かる．各ノード下の話者の音素分布間距離行列の平均行列を使えば，各ノードに対する日本人英語の構造が得られる．今後，あるノード下の学生群とそのノードの構造とが本当に合致しているかどうかについて，英語教師と実際に議論していく予定である．

なお，音素群が成す構造に影響を与える要因としては，読み上げ文以外にも，発話速度，調音努力，方言なども考えられる．音声収録の際にはこれら要因は統制しなかったため，各実験結果には上記要因による雑音が混入していることは否めない．

4. 同一話者による異なる習熟度の英語発音を用いた分析

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，そのような発音のための指導も試行錯誤的ながら行なってきた．一方，日本人である以上，日本語訛りの英語を発声することも容易である．そこで，筆者が発声した通常の英語と故意に日本語訛りで発声した英語を使い，興味深い実験を行なった．**表 2** に用意した異なる発音を示す．USA/F12 は女性の米語母語話者であり，英語教師である．これと筆者の 2 種類の英語である．以降，米語母語話者の英語を発音 F，筆者の通常の英語を発音 A，故意に訛らせた英語を発音 B とする．さて，A が教育的に B よりも F に近いと判定されるべきであるとするなら，この条件が音声認識技術にとって一番困難な条件であることは自明である．何故なら，F と B は習熟度以外は全て mismatch，A と B は習熟度以外は全て match としてあるからである．

各発音ともサブセット 6 の 60 文音声が用意された．また，比較対照として，USA/M08（男性，発音 M）の 60 文も使用した．F/M/B から HMM を作成し，以下に示す方法で判定した．

表 2　実験で使用した三種類の英語発音
Table 2　Three kinds of pronunciations used in the experiment

話者	USA/F12(F)	筆者 (A)	筆者 (B)
性別	女	男	男
年齢	約 50	36	36
マイク	Sennheiser	特価品	特価品
録音室	防音室	リビング	リビング
AD	SONY DAT	PowerBookG4	PowerBookG4
習熟度	perfect	good	Japanized

- F/M/B モデルと音声 A 間の尤度スコア $P(o|M)$
- F/M/B モデルと音声 A 間の事後確率スコア $P(M|o)$
- A からもモデルを構築し，算出した構造歪みスコア

事後確率スコアは，モデルと入力話者間の相性・整合性を正規化する目的で発音評定では広く使われている．

$P(o|M)$ 及び $P(M|o)$ 使用時の結果を**図 8**（上 2 つが $P(o|M)$，下 2 つが $P(M|o)$）に示す．各々のスコアを元に A の存在位置を内分点として示している．$P(o|M)$ 使用時は，A は限りなく B に近い．これは発声者が同一話者である以上（日本語・英語を知らない人の判断を考えれば），至極当然である．一方 $P(M|o)$ であるが，事後確率の近似式を考えると，本スコアは入力話者とモデルとの相性を正規化する働きを持つが，結果は「異なる教師が全く異なる採点を下した」形になっている．音声認識技術の不安定性が図らずも露呈してしまった．さて，構造歪みによる結果を**図 9** に示す．上 2 つが A の位置を示しており，教師に因らず同じ評価を得ている．下 2 つは，サブセット 6 を読み上げた米語話者，日本人全て同一軸上に示したものである．軸の上下で女男を示している．筆者は米語話者と日本人の境界に位置している．米語話者の中にいる日本人は 202 人中唯一のバイリンガル話者である．また，図より分かるように，「バイリンガル以外の全ての日本人が筆者以上に，筆者（B）に近い」と判定されている．このような現象は，単純なスペクトル照合では不可能である．本提案手法が，発音習熟度のみに着眼し，他の静的歪みに関する要因を完全にそぎ落としていることを明確に示す結果である．なお，構造に着眼せず，話者適応技術を用いても同様の結果が得られると考えられるが，適応技術を使うと日本人英語への適応がかかり，1) 評定スコアが不必要に，より高くなる危険性がある，2) 適応データ量に依存した方法論となる，などの問題がある．本手法は適応するのではなく，評定に不必要な要因を表現する次元を持たない音声表象に基づく検討を行なっており，このような問題は一切生じない．

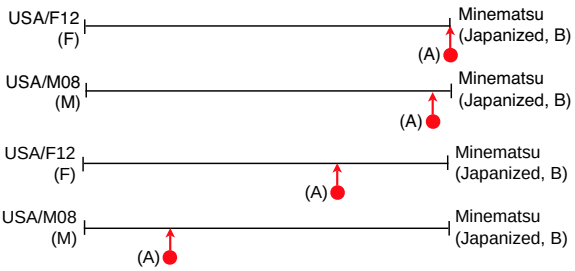

図 8　$P(o|M)$ と $P(M|o)$ による発音評定
Fig. 8　Proficiency rating with $P(o|M)$ and $P(M|o)$

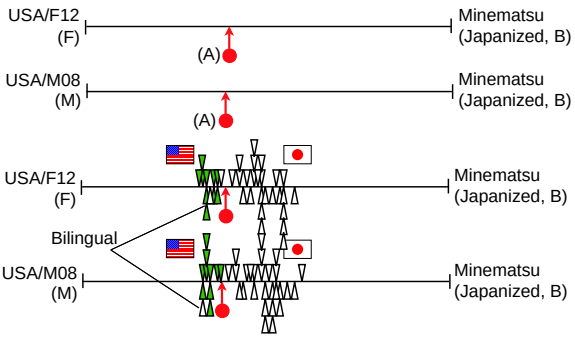

図 9　構造歪みスコアを用いた発音評定
Fig. 9　Proficiency rating with structural distortion score

5. ま と め

発音習熟度の評定とは一切無関係の種々の音響要因を消失させた形で音声を「構造として」表現し，その構造的歪みに基づいた自動発音評定が可能であることを示した．また，その消失が完全であることを実験的に示した．しかし，検討した手法はあくまでも評定値を算出するだけであり，学習者に対して具体的にどのような訓練をすべきか，という教示をするには至っていない．文献 [12] で母語話者音声との音響的照合を全く使わずに効率的な教示生成を行なっているので，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] http://www.mext.go.jp/b_menu/shingi/chousa/shotou/020/sesaku/020702.htm
[8] 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)
[9] 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)
[10] A. Neri, *et al.*, “Automatic speech recognition for second language learning: How and why it actually works,” Proc. ICPhS, pp.1157–1160 (2003)
[11] 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資料，SP2003-179 (2004)
[12] 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)
[13] C. J. Leggetter *et al.*, “Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models,” Computer Speech and Language, vol.9, pp.171–185 (1995)
[14] T. Anastasakos, *et al.*, “A compact model for speaker-adaptive training,” Proc. ICSLP’96, vol.2, pp.1137–1140 (1996)