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

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　人間間の音声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，この歪みに一切影響を受けない音声の物理的表象が峯松により提案されている．この音声表象では，音声事象を確率論的に有限個の状態として捉え，その状態群を情報論的に構造として捉えることで実現される．この表象を用いて学習者の発音を表現すると，上記の静的歪みを表現する次元を一切持たないため，主に，外国語発音における母国語依存性のみが表現される．外国語学習における発音の評定基準として「母語話者らしい発音」ではなく,「明瞭な（伝わる）発音」の必要性が叫ばれている．本研究では「明瞭な発音」を認知科学的に「心的辞書アクセスに必要な処理量が少ない発音」と定義する．そして，各学習者から抽出される発音の音響的構造を，対象言語の語彙が持つ言語的構造と比較・照合することで，心的辞書アクセスに必要な処理量を推定する．また，その処理量削減に効果的な学習指針についても学習者毎に推定可能であることを示す．
キーワード　音声の普遍構造，発音学習，明瞭度，コホートモデル，学習指針

Estimation of intelligibility of the pronunciation based on compatibility of acoustic and linguistic 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 states based on probability theory and the states are viewed as structure based on information theory, the structure cannot be changed by the above distortions. When a language learner is modeled by this new representation of speech, since the model has completely no dimensions for the static distortions, only the dependency of the learner's pronunciation on his/her mother tongue is seen in the model. Nowadays, goal of pronunciation training has been shifted from "native-sounding" to "intelligible" pronunciation. In this work, the more intelligible pronunciation is defined as the pronunciation where mental lexical access is possible with lower cognitive load. By comparing the acoustic structure in a learner's pronunciation and the linguistic structure in the target language's vocabulary, the expected cognitive load is calculated for individual learners. Instructions are also automatically generated to reduce the cognitive load efficiently with the minimum training efforts.
Key words　universal structure, pronunciation training, intelligibility, Cohort Model, instructions

1. はじめに

近年の音声情報処理技術の進展に伴い，語学学習に対する技術的サポートが議論されるようになった．既に多くの音声研究者，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 [1]，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 [2].

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．日本における英語教育では，かつて「ネイティブ神話」という言葉で表される発音教育指針があった．即ち，母語話者のような発音を是とする教育方針である．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．舞台の上では，母語話者よりも母語話者らしい発音を求められ,「ネイティブ神話」のみの発音訓練となる．しかし，一般の語学学習者が全て舞台役者を目指す訳ではなく，また，実践的な目標として掲げるべきは「通じる英語・明瞭な英語」である．事実英語教育の現場でも，明瞭な英語が発音学習の目標として掲げられるようになってきている [3]．CALL システムの研究が広く行なわれているが,「母語話者のような」発音を是とするのではなく,「明瞭な」発音を明示的に是としたシステム開発を筆者は知らない．これは，2 種類の発音に対して，(母語話者音声との音響的類似度に頼らずに）どちらがより明瞭かを判断する手段が存在しないことが原因である．母語話者音声をモデル音声としない発音評定を目指す場合は,「日本語訛りは残っているが，社会的に何ら問題が無い日本人英語」をモデル音声とすることで解決するが，この場合，訛り方が発声者によって多種多様であること,「社会的に問題が無い」という事象の定量化が困難であるなどの問題が残る．

「母語話者のような」発音を是とするシステム開発しかできないのは何故か？CALL システムを現在の音声認識（音響照合）技術の一アプリケーションとして見た場合，自ずと入力音声と音響モデルとの照合となる．上記した様に日本人英語音声による音響モデルとの照合が問題を有する場合，結局母語話者音響モデルと学習話者音声間の音響的照合へと落ちる．結局，現在の音声認識の枠組みで発音教育を支援する限り，それは「母語話者のような」発音を是としたシステム開発にならざるを得ない．そしてそれは,「ネイティブ神話をいつまで唱えるのか？」という批判に常に曝されるシステムになる．なお，筆者は「ネイティブ神話」を否定している訳ではない．これはあくまでも筆者の経験であるが，正しい英語調音を円滑に行なうために必要な筋肉をつけること，及び，正しい英語リズムを生成するための呼吸法を体得することで，英語特有の音変化が至極当然の結果として（即ち，そのように変化させる方が「楽な」発声であるとの感覚に落ちる）受け入れられるようになり，その結果，ヒアリングの能力も向上する，と考えているからである．換言すれば，英語特有の音変化を頭で理解するのではなく，筋肉で理解することができると考えているからである．事実，昨今の英語教育で（特に学生に）嫌われる傾向にある「母語話者のような発音」と「語彙力の向上」に対して，これらが最終的には英語能力向上の一番の近道であるとの観点からの教育論も展開されている [4]．つまり「英語調音のための適切な筋肉・呼吸法を体得すること」及び「単語を覚えるのではなく，個々の単語の源となる語源に対する理解を深めること」の追及である．

結局「学習者が英語教育に何を求めるのか」の一言に尽きる．海外旅行でホテルにチェックインできるための英語能力から，公の場での英語による激論に耐えうる英語能力まで，学習者の要求は千差万別である．彼らの要求に対して適切な目標を設定し,「現在その目標からどのくらい離れているのか（つまり学習者は現在どのような状況にあるのか)」「その状況にある学習者にとって最も効率良く目標に近づく訓練はどれか」「同等の他学習者の学習履歴を元に推定される当該訓練の効果」といった情報を提供する枠組みが求められるべきだと筆者は考えている．

本研究では「明瞭な」発音を是とする教育方針に立脚し，母語話者音声との音響的照合を一切用いずに，1) 発音の自動評定が可能であること，2) その学習者が次に行なうべき発音訓練の指針に関する教示が可能であること，を示す．本研究において学習者発音と比較する対象は母語話者音声ではなく，英語という言語そのものである．学習者の現在の発音に基づいて英語という言語を運用する場合に，その発音は「英語という言語にとってどの程度都合が良いのか」を定量化する．なお，本研究では考察対象としていないが，世界中の外国語訛りの英語発音，及び母語話者による英語発音を比較した場合，母語話者発音がどの外国語訛りの発音よりも，英語という言語にとって「都合が良い」発音であるならば，究極の英語能力を求める学習者に与えるべき教育指針は「ネイティブ神話」となる．本研究の遂行には，従来の音声認識技術に基づく音声表象は不必要であり，最近峯松により提案されている「音声の音響的普遍構造 [5], [6]」に基づく音声表象を用いる．以下，その説明から行なう．

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．

(注1)：平方根を使う理由は文献 [6] を参照して戴きたい．

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き，音素分布群が成す構造は不変となる．

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

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，ユークリッド空間における構造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 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，ベクトル b 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，点の集合である構造に対して，各点が平均ベクトルであり，各点に分布が付随している状況を考える．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共通のアフィン変換をかけた場合，任意の二点間距離は不変である．上記の変形の中でこれを満たす変形は回転とシフトのみである．つまり，行列 A による変形は構造の回転として観測され，b による変形は構造のシフトとして観測されることになる．

b を足す演算は，伝達関数を一つ掛ける演算であることを考えると，これは，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平均値のモデル化であることを考えれば，話者性の一部も b となる．では 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近似される．文献 [7] 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導いている（逆は真ではない）．結局声道長の違いによる音響的差異は構造の回転として解釈できる．聴覚系の周波数特性であるバーク尺度も周波数ウォーピングである．この場合，音響刺激が変形するのではなく，観測系の座標軸の非線形伸縮であり，これは変数変換すれば，線形軸に対して音響刺激構造が回転することと同値である．以上の議論より，音声の生成・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の歪みは，音声を構造化する本表象では原理的に歪みになり得ない．

提案する音声表象で語学学習者を表現した場合，主に「外国語発音の母国語依存性のみ」が観測されることになる [5]，また，性別，年齢，録音環境特性などの要因が完全にそぎ落とされている様子も実験的に確認されている [6]．なお，本音声表象の視覚化は，樹型図や多次元尺度法などによって可能となる．

3. 音響的普遍構造と言語的普遍構造との整合性に基づく発音評定

「一切の静的歪みに無縁の音声の物理表象」が存在することが峯松により示された（音声の音響的普遍構造）．この構造は学習者毎に異なってくる．どのような構造が英語という言語にとって都合の良い構造と言えるのだろうか？この問いを音声知覚モデルを参照しながら解くことを考える．

3.1 音声知覚モデルに基づく整合性の定量化

音声知覚モデルとして孤立単語音声の知覚モデルを考える．種々のモデルが提案されているが，何れも，心的辞書内の単語が音響的・言語的刺激によって活性化され，活性化された単語数がやがて 1 つとなって知覚過程は終了する，という前提を置いているものが多い．つまり，知覚が完了する以前は活性化単語が複数存在している．ここで「明瞭な」発音を「知覚の途中において活性化する単語数がより少ない発音」と定義する．日本語音素数が約 25，英語音素数が約 40 である事実を考えると，日本語音で英語を発音すれば，自ずと 1 対多のマッピングとなり，何らかの音素混同が生じる．結局「知覚途中における活性化単語数」は増えることになる．この語彙密度の増量でもって学習者発音が呈する構造を評価する．例えば日本人英語の場合，/s/と/th/の混同，/r/と/l/の混同が代表例としてあるが，英語という語彙体系を考慮した場合，どちらの混同がよりダメージが大きいのかを定量的に議論することに相当する．なお，本研究ではコホートモデルを前提として語彙密度を計算する．

3.2 コホートモデル

オリジナルのコホートモデル [8] では，単語音声知覚過程を left-to-right 処理系として構成する．例えば，語頭音が/s/である場合，語頭に/s/を持つ単語セットが心的辞書内で形成される（活性化される）．音声入力が継続され/str/となった場合，語頭に/str/を持つ単語セットが形成されると共に，上位情報（言語情報）と整合のとれない単語は除かれる，としている．単語同定過程の途中で構成される活性化単語セットがコホートであり，コホートサイズが 1 となった時点で単語知覚は完了する．コホートモデルでは，語頭の音素の識別に失敗した場合，単語の同定に失敗することになる．つまり，語頭の音素同定に非常に高い精度を要求するなどの問題点も存在するが，単語知覚のモデル化を考える場合，頻繁に参照されるモデルである．なお，整合性算出におけるコホートモデルの必然性は無い．本モデルは非常に単純であり，その実装が簡単に行なえるだけである．

本研究で用いる音声の音響的普遍構造に基づく音声表象は，音韻論をベースとした音声の捉え方である [5][(注2)]．音韻論とは，対象とする言語音（音素）に対して，音素セットに内在する規

(注2)：但し，厳密には音韻論とも音声学とも異なる音声表象である．

則・関係，及び，音素の並びに内在する規則・関係を記述する科学であり，音声の音響的普遍構造は前者の音韻論に端を発している [5]．さて，コホートモデルを単語知覚モデルではなく，対象言語の語彙体系のモデルとして考えた場合，これは，音声認識で広く使われる，音素を単位とした木構造辞書となる．そこには，英語として現れない音素並びは一切含まれない．つまりこの木構造辞書は，音韻論における「音素並びに内在する規則・関係」のモデルとして考えることができる．結局本研究は，音韻論で議論する二種類の構造間の整合性を定量化することに相当する．前者の構造は文献 [5] で定義した音響的普遍構造であり，学習話者を特定すれば定まる．一方後者の構造は言語を特定すれば，その言語の母語話者が共通して持つ言語感（の一側面）であり，本研究ではこれを言語的普遍構造と呼ぶ．この二種類の普遍構造間の整合性を音声知覚モデルの上で検討する．

3.3 シラブルを単位としたコホートサイズの推定

各学習者の音響的普遍構造に対する，第一シラブル入力時のコホートサイズを推定する．一般にコホートモデルは音素を単位とすることが多いが，本論文では，知覚単位，発声単位に基づいたコホートを考える．即ち，英語の知覚単位の一つであり，同時に発声単位であるシラブルを単位としてコホート構成を考える．語彙セットとしては，語彙数 20K の WSJ の unigram を使用した．この辞書の全エントリーを tsylb [10] を用いてシラブルに分割した．なお，音響的普遍構造は 60 文発声から求められるため二重母音は考慮していない．そのため，二重母音を第一シラブルに持つ単語は無視した．その結果，語頭に位置する異なりシラブル数は約 3,200 種類であった．

各異なりシラブルについて $CS_0(s_i,\theta)$ を求める．$CS_0(s_i,\theta)$ とは，シラブル s_i 或は s_i から音響的に近い（則ち，距離 θ 以下となる）シラブルを語頭に持つ単語数の総和である．次に $CS_1(w_j,\theta)$ を求める．$CS_1(w_j,\theta)$ とは w_j の先頭シラブルを $s^1(w_j)$ とした時 $CS_0(s^1(w_j),\theta)$ である．つまり w_j の先頭シラブル或はそのシラブルから音響的に近いシラブルを先頭とする単語数である．最終的にコホートサイズの全単語に対する期待値，ECS(Expected Cohort Size) を次式によって求める．

$$ECS(\theta) = \sum_j p(w_j) CS_1(w_j,\theta) \qquad (6)$$

$p(w_j)$ は w_j の unigram 値である．こうして，単語の生起頻度まで考慮したコホートサイズが，シラブル間距離閾値 θ の関数として定義される．単語の生起頻度を考慮せず，また，日本人英語・米語の不特定話者モデルを用いたコホートサイズ推定については，文献 [9] を参照して欲しい．任意の 2 シラブル間の距離は，音素（或は音素状態）間距離行列を参照し，シラブル（音素連鎖）間の DP により計算できる．即ち，本研究で検討する音響的普遍構造と言語的普遍構造の整合性の定量化は，音声を構造として表象する際に得られるパラメータ（音素間，或は音素状態間距離行列）だけが音響情報として必要となる．

3.4 結果と考察

表 1 に実験条件を示す．ERJ データベース [2] の各話者（全 222 名）に対して GMM を用いて音素を音響的にモデル化し，

表 1 実験条件
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	MFCC(1〜12)+ΔMFCC+ΔPower
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	3 状態 1 分布（GMM）
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

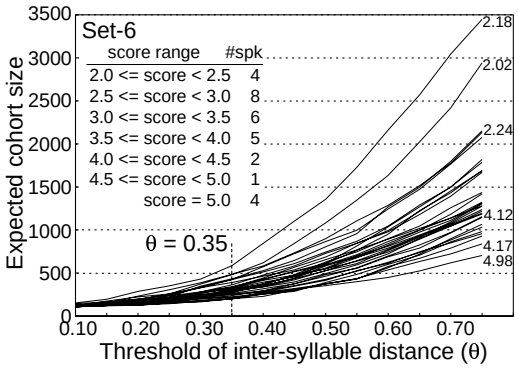

図 1 シラブル間距離 θ の関数として表現した ECS
Fig. 1 ECS as function of inter-syllable distance θ

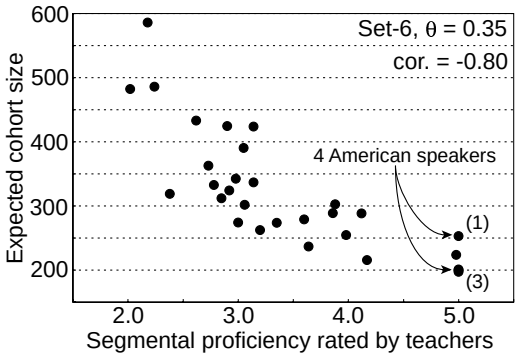

図 2 自動発音評定結果
Fig. 2 Result of pronunciation proficiency estimation

音素間距離行列を算出した．なお，構造の回転を考える場合，分散共分散行列は全角でなければならないが，本実験は対角行列で行なっている．以下の実験結果に，このような実験条件の不具合の影響があることは否めないことを断っておく．

図 1 にシラブル間距離 θ の関数として ECS を示した．音響的普遍構造は文セットに依存する傾向があるので，ここでは比較的幅広い発音習熟度の話者が得られた（ERJ データベースには英語教師による採点結果も含まれている）文セット 6 の話者のみを使用している．図中，採点スコアが 5 点満点の話者は母語話者（4 人）である．図右側には，本セットを読み上げた日本人の中で英語教師の採点の上位 3 人，下位 3 人のスコア及び位置を示している．成績下位者の方がコホートサイズが非常に大きいことが分かる．成績上位者の一部が母語話者よりもコホートサイズが小さく見積もられている．音響的普遍構造は音

素間距離のみに基づいて音声を構造化しているため，話速や調音努力など音素間距離を変動する要因の影響を受ける．例えば，母語話者であっても話速が早い場合は音素間距離が小さくなり，構造が小さくなる傾向がある．スコア逆転の要因の一つとして考えている．収録時に発声速度や発話スタイルを統制するか，事後的に正規化する処理が必要である．

図 1 に対して，$\theta = 0.35$ において ECS を求め，それと英語教師による評定結果との関係を**図 2** に示す．なお，母語話者は 5 点満点としている．相関係数は -0.80 であり，比較的良好な対応がとれている．注意して戴きたいのは本発音評定を行なう際に，母語話者音響モデルとの照合のみならず，母語話者音声データそのものを一切用いていない点である．従来の CALL システムは全て，母語話者音響モデル・音声データとの比較に基づく方法論であった．その場合常に「ネイティブ神話をいつまで唱えるのか？」という批判を受けることとなる．本評定手法は「外国語学習における発音習得は，その言語に内在する言語的構造と整合性の高い音響的構造を口に宿すことである」という立場，及び「より明瞭な発音とは，聞き手が心的辞書検索をする場合に，その検索タスク量がより小さくなる発音である」という立場に基づいた全く新しい発音評定方法である．なお，音響的普遍構造に基づいた音声表象しか必要としないため，音声認識の世界で嫌われる mismatch 問題が原理的に発生しない．

4. 音響的普遍構造と言語的普遍構造との整合性に基づく効率的な学習計画の自動生成

4.1 音響的普遍構造間における部分的構造の置換

本研究では，音声の音響的普遍構造推定に基づく音声表象を用いている．本表象では，音声生成・収録・伝送・再生・聴取の過程で不可避的に混入する乗算性歪み，線形変換性歪みに一切依存しないため，性別，年齢，体格，録音環境といった条件を完全にそぎ落とすことが可能である [6]．このそぎ落としによって，スペクトルや波形といった従来の音声表象では完全に不可能である演算が可能となる．それは異なる二表象間の部分的な置換である．例えば教師（1 名）音声から得られる構造の一部を，学習者の構造の埋め込むことを考えると，これは，その構造の取得前・後の学生の様子を模擬することになる．このような演算をスペクトルや波形で行なえば，当然性別・年齢などの情報までも置換することとなり，無意味な演算となる．音声を構造化することによって初めて可能となる演算の一例である．

何故置換するのか？それは，コホートサイズを最小化する部分置換を求めれば，その部分置換を実現する学習が，その学習者を最も効率良く，目標とする構造へと導くからである．

4.2 コホートサイズ最小化基準に基づく部分構造の選択

音素間距離行列を $C = \{c_{ij}\}(1 \leqq i, j \leqq M)$ とする．「音素 p に関する関係」を $\{c_{pj}\}$ 及び $\{c_{ip}\}(1 \leqq i, j \leqq M)$ で定義する．即ち，要素 $c_{pp}(\equiv 0.0)$ の上下・左右に位置する要素を「音素 p に関する関係」とする．当然音素の種類数（M）だけこの関係は存在するが，どの関係の置換がコホートサイズを最も低減するのかを検討する．音素 p_0 の関係の置換が最もコホートサイズを低減する場合，次の学習対象を音素 p_0 と考える訳である．

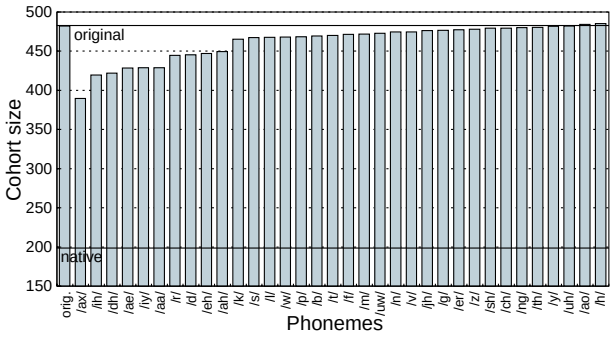

図 3　一音素関係の置換によるコホートサイズの減少

Fig. 3　Cohort size reduction by a single replacement of phonemic relation

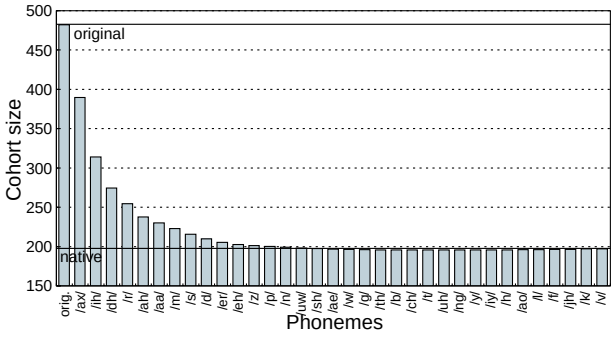

図 4　逐次的な音素関係の置換によるコホートサイズの減少

Fig. 4　Cohort size reduction by sequential replacements of phonemic relation

ERJ データベース中の話者 RYU/F06（分節的側面に着眼した英語教師の採点は 2.02）を用いてコホートサイズを最も低減する音素関係を求めた．なお，置換対象としたのは母語話者 USA/F08 である．まず 1 回の音素関係置換で最大のコホートサイズ低減をもたらす音素について検討した．結果を**図 3** に示す．**第 3.3 節**で示したコホートサイズ推定を行なったところ（$\theta = 0.35$），この話者のオリジナルコホートサイズ（置換前）は約 480 であった．図より第一の学習対象音素は/ax/（弱母音，schwa）であることが分かる．/ax/の関係が是正された場合，次なる学習対象音素は，/ax/関係置換に対する事後的な操作によって求まる．**図 4** に RYU/F06 に対して得られた，最も効率良く音響的普遍構造を USA/F08 へと近付けるための音素学習順序について示す．Schwa 以外にも，日本人にとって正しい調音が比較的難しい音素に対して優先度が高く見積もられていることが分かる．本分析を他の学習者に対して行なったところ，例えば schwa 音の置換が最も優先順位が低く見積もられる学習者がいることが分かった．本研究で検討した部分置換は，音素 p に関する関係を一度に全て置換してしまう．これでは，p とどの音素との関係が最も劣悪なのか，といった情報が欠落してしまうため，非常に粗い分析をしていると考察される．部分構造として全ての音素対を個別に考え（音素対数は ${}_MC_2$），コホートサイズを最も低減させる上位 N 個の音素対を算出し，そこに頻出する音素を次なる学習ターゲットとするなどして分析の精度を上げることで解決できると考えている．

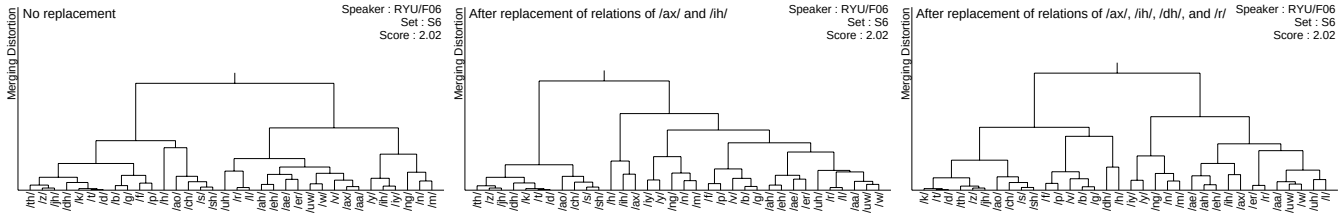

図 5　学習者・教師間の音響的普遍構造間の部分置換に基づいて予測される学習者発音構造の変化の様子

Fig. 5　Prediction of change of a learner's pronunciation structure via. replacing sub-structures between the learner and a teacher

なお，構造の部分置換に基づく学習指針の教示は非常に興味深い語学教材を可能とする．教師と学習者の一対一の間で部分構造を置換する訳だが，この時，教師，学習者共に音響的普遍構造に基づく表象が使われるため，性別，年齢，体格，収録環境といった要因は一切そぎ落とされる．その結果，学習者が教師を選ぶことが可能となる．教師音声としては，学習者が読み上げた文と同一の文の音声が必要となるが，これは HMM 合成を用いることで可能となる(注3)．結局数百〜千文ほどの英文を教師から提供してもらえれば，任意文の発声に対応できることになる．もし映画俳優や歌手など，学習者の動機を高める「個性ある」教師の音声を利用することができれば，学習者の「憧れ」と学習者とを一部置換しながら，学習者の発音を「憧れ」へと近付ける最短パスを提示する発音教材が可能となる．

4.3　訓練後の学習者が呈する音響的普遍構造の予測

音響的普遍構造の部分置換に基づいて，学習対象音素の優先順位が推定できることを示した．部分置換によって学習者が呈する構造は変化する．構造が変化すれば，それを視覚化した(例えば) 樹型図も変化する．つまり，訓練によって学習者発音がどのように変化すると予想されるのか，を学習者に提示できることになる．**図 5** に置換前の発音構造，及び**図 4** に示した学習音素順位の第 2 位，4 位までの音素関係を置換した時に予想される学習者の発音構造を示す．このような教示は，学習者の動機を高める意味において重要であると考えている．なお，本節で議論している学習の優先度や，訓練後の学習者の予測はいずれも技術的な可能性の議論に留まっており，英語教師との議論や実際の教材としての効果に関しては今後の課題である．

「構造の部分置換に基づき，学習者構造を対象言語に対して，より都合のよい構造へ変化させる」方法論は，遺伝的アルゴリズムにおいて「アルゴリズムを構造化し，その部分構造を変化させることでアルゴリズムを変化させる」という方法論に類似している．その意味において**図 5** は学習者の進化の過程のシミュレーションとして位置づけることもできる．

5. ま と め

音声の音響的普遍構造に基づいて学習者を表現し，この構造と，対象言語の語彙が持つ言語的普遍構造とを音声知覚モデルの上で比較することで，学習者の発音が，その言語にとってどの程度「都合が良い」のかを算出可能であることを示した．そしてそれが「母語話者音声を必要としない」全く新しい発音評定技術となることを示した．これは「明瞭な」発音を是とする近年の発音教育に根差した方法論である．更に，教師・学習者間で発音構造を部分的に置換することで，効率的な学習計画を立案することが可能であり，かつ，各発音訓練による学習者の進化の様子さえも予測可能であることを示した．

本研究や文献 [6] で提案した方法論の幾つかは，いずれも，音声学という「個々の音を眺める」科学に立脚する現在の音声工学では凡そ不可能である．文献 [5] で述べたように，音響的普遍構造は「音声学を捨てる」という発想の下で,「音韻論」での議論を物理の上で再構築することで生まれた全く新しい音声の物理的表象である．その意味において，従来の音声工学では不可能だったことがいとも簡単に実現されて何ら不思議ではない．現在「音響的普遍構造を満たす音セットを調音合成器を用いて探索させた場合，対象言語の個々の音が導かれるのか」という問いに答えるべく実験の準備を進めている．即ち「調音器官の構造的制約の下で音響的普遍構造の実現を図った場合，それは，個々の言語音の音響的特性と調音的特性を導き出すのか」という問いかけである．もし個々の音が正しく導かれた場合「音響的普遍構造＋調音の構造的制約」は，音響音声学と調音音声学(現在の音声工学の基盤を成す科学) を副次的に導出可能であることを実験的に示したことになる．そして「正しい構造」は「正しい発音」の十分条件として位置づけられることになる．

(注3)：HMM 合成をスペクトル系列生成器としてのみ使用する．

文　献

[1] 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)
[2] 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)
[3] D. Crystal, English as a global language, Cambridge University Press, New York (1995)
[4] http://www.scn-net.ne.jp/~language/
[5] 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)
[6] 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)
[7] M. Pitz *et al.*, “Vocal tract normalization as linear transformation of MFCC,” Proc. EUROSPEECH, pp.1445–1448 (2003)
[8] W. D. Marslen-Wilson, *et al.*, “The temporal structure of spoken language understanding,” Cognition, vol.8, pp.1–71 (1980)
[9] 峯松他，“米語における音素体系及び語彙体系に着眼した日本人英語の発声と聴取に関するコーパス統計分析”，音声研究，vol.7, no.3，pp.77–91 (2003)
[10] Tsylb: http://www.nist.gov/speech/tools/tsylb2-11tarZ.htm