学習者が呈する音韻構造と対象言語の語彙構造との整合性に基づく発音明瞭度の自動推定

峯松 信明（東京大学大学院情報理工学系研究科）
mine@gavo.t.u-tokyo.ac.jp

1　はじめに

近年の音声情報処理技術の進展に伴い，CALL システムの研究開発が広く行なわれるようになった (1; 2)。発音教育の技術的支援を考える場合，学習者の音声を母語話者のそれと比較・照合し，スコアを提示する形態のものが多い。即ち母語話者からの差異を定量化する。しかし，周知のように英語は国際語としての位置を確立し，発音教育に求められるべきものは「母語話者のような発音」ではなく「通じる・intelligible な発音」であるとの声を聞くのも事実である。前者ではなく，後者の発音評価を技術的に求めた場合,「母語話者からの差異」に基づく方法論では不十分であることは自明である。

「通じる英語」とはどんな英語だろうか？聴取者が日本人であれば，日本人英語が最も通じる英語である。即ち「通じる英語」は聴取者の母国語に依存し，厳密には聴取者そのものに依存する。その結果，地球上の全人類を考えれば，60 億種類の通じる英語の定義があると考えることもできる。このような定義が雑然とした対象を学習目標に位置付けざるを得ないのだろうか？

本研究は，聴取者ではなく，学習対象言語（米語，英語など）に依存する形で，学習者発音の明瞭度を推定する方法を提案する。と同時に，どの音素の発音から修正するのが最も効率良く明瞭度を向上させるのか，その順序の推定が可能であることも示す。地球上には約 20 億の英語学習者が存在する。彼らが一人一人異なるとすれば，20 億種類の異なる訛った英語が存在する。明瞭度向上のために 20 億種類の異なる教示が必要となる訳だが，それを可能にする技術的枠組みの提案である。

本研究では，音声から「話者・音響機器の違いなどの非言語情報を表現する次元を消滅させることで」定義される音声の物理表象である音響的普遍構造 (3) を用いる。次節で簡単に説明する。

2　音声に内在する音響的普遍構造

音響空間中に，各音素が位置している様子を考える。構造音韻論では，音素群が全体として持つ構造を議論する。例えばヤコブソンはフランス語音素に対して，弁別素性を用い，幾何学的な構造を提案している。音素毎に音響モデルを構築し，モデル群を構造化・相対化すると種々の物理情報が消失される。この消失される情報について検討する。なお，n 種類の音素が存在した場合，その n 音素構造を規定するためには，${}_nC_2$ だけ存在する全ての対角線・辺の長さがあればよい（距離行列）。

各音素を，ケプストラムで構成される多次元ガウス分布であると仮定し，音素間距離をバタチャリヤ距離の平方根で定義する。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。

$$BD(P_u,P_v) = -\ln\int_{-\infty}^{\infty}\sqrt{P_u(x)P_v(x)}dx = \frac{1}{8}\mu_{uv}^T\left(\frac{\Sigma_u+\Sigma_v}{2}\right)^{-1}\mu_{uv}+\frac{1}{2}\ln\frac{|(\Sigma_u+\Sigma_v)/2|}{|\Sigma_u|^{\frac{1}{2}}|\Sigma_v|^{\frac{1}{2}}} \quad (1)$$

μ_u は u の平均ベクトル， μ_{uv} は $\mu_u-\mu_v$ を, Σ_u は u の分散共分散行列である。二つの確率密度の相乗平均に対する全領域の積分値（0.0 以上 1.0 以下。即ち確率としての意味付けが可能）の対数として距離を定義している。積分値を確率値と考えれば，これは自己情報量である。結局，音素を分布として捉え，分布間距離を情報理論的に規定することで定義される構造（距離行列）を考える。

バタチャリヤ距離の持つ性質として一次変換不変性がある。二つの分布に対して，同一の一次変換（$c'=Ac+b$）を施したとしても，分布間距離は変らない。即ち任意の一次変換によって距離行列（即ち構造）は不変である。なお，音素を点として考えた場合，一次変換によって一般的に構造は歪んでくる（アフィン変換）。ケプストラムに対する一次変換が持つ音声学的意味を考える。音声認識・合成の分野で話者性やマイク特性を変更するための操作が広く行なわれるが，行列 A をかける演算は，

表 1: 実験で使用した三種類の英語発音

話者	USA/F12(F)	話者 NM(A)	話者 NM(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

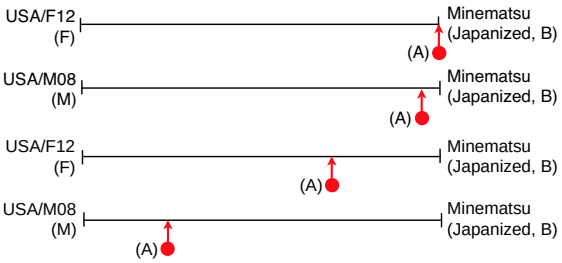

図 1: $P(o|M)$ と $P(M|o)$ による発音評定

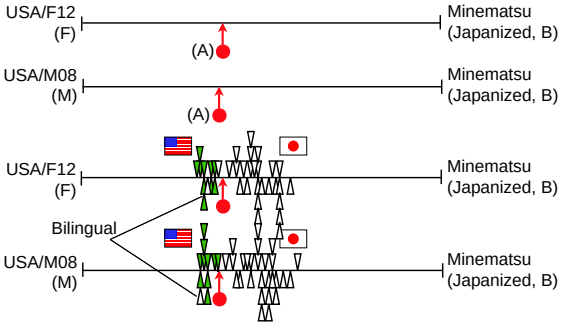

図 2: 構造歪みスコアを用いた発音評定

話者や聴取者の生理学特性の差異，ベクトル b を足す演算は音響機器特性及び一部の話者特性の差異を表現する。即ち，ケプストラムの一次変換は，音声に不可避的に混入する非言語情報の最も単純なモデルとして考えられる。そして，この変換によって構造は一切変化しない。即ち本構造は，構造音韻論で議論される音素群構造の物理実装として定義され，音声の音響的普遍構造と呼ばれている (3)。次節でまず，構造として表象された学習者発音における，非言語情報の消失を実験的に示す。

3　同一話者による異なる習熟度の英語発音を用いた分析

英語劇における発音指導経験者（日本人話者 NM）より，通常の英語と故意に日本語訛りで発声した英語を収録した。**表 1** に用意した資料を示す。USA/F12 は女性の米語母語話者である（英語教師）。これと話者 NM の 2 種類の英語である。以降，米語母語話者の英語を F，話者 NM の通常の英語を A，故意に訛らせた英語を B とする。さて，教育的に，A が B よりも F に近いと判定されるべきであるならば，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，A と F は習熟度以外は全て「不一致」，A と B は習熟度以外は全て「一致」としてあるからである。

各発音とも同一の 60 文音声が用意された。また，比較対象として，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。

- 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（具体的な定義については文献 (4) 参照）

事後確率スコアは，モデルと入力話者間の整合性（相性）を正規化する目的で広く使われている。

$P(o|M)$ 及び $P(M|o)$ 使用時の結果を**図 1**（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 2** 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，同一セット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。話者 NM 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 ERJ 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話者 NM 以上に，話者 NM（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。音声の構造化が，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。

4　学習者が呈する音韻構造と対象言語の語彙構造との整合性に基づく発音評定

4.1　音声知覚モデルに基づく発音の明瞭度の定量化

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。種々のモデルがあるが，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，活性化単語数がやがて 1 つとなって知覚は終了する，という前提を置いている。つまり，知覚完了以前は，活性化単語が複数存在する。ここで「明瞭な発音」を「知覚途中において活性化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，日本語英語は，自ずと 1 対 N 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/の混同が代表例であるが，英語という語彙体系を考慮した場合，どちらの混同がダメージが大きいのかを定量的に論じることに相当する。なお，本研究ではコホートモデルを使う。

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，構造音韻論の物理実装という側面を持つ (3)。この音素群に観測される構造と，対象言語の単語に見られる音素並びに見られる構造（コホートモデルによる単語知覚は，単語群を例えば音素を単位とした木構造辞書として考えることと等しい）との間に観測される整合性を定量化することに相当する。学習者を特定すれば，音素構造が決定され，学習対象方言（米語・英語など）を特定すれば語彙構造が決定される。この異なるレベルの言語学的構造の整合性を，定量的に，音声知覚モデルの上で検討する。

4.2　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(5) 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。

$$ECS(\theta)=\sum_j p(w_j)CS_1(w_j,\theta) \tag{2}$$

$p(w_j)$ 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。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 DTW により計算できる。即ち，本研究で検討する音素構造と語彙構造の整合性（相性）の定量化は，音素群を構造として表象する際に抽出されるパラメータ（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。

表 2: 実験条件

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（PRONLEX 表記）

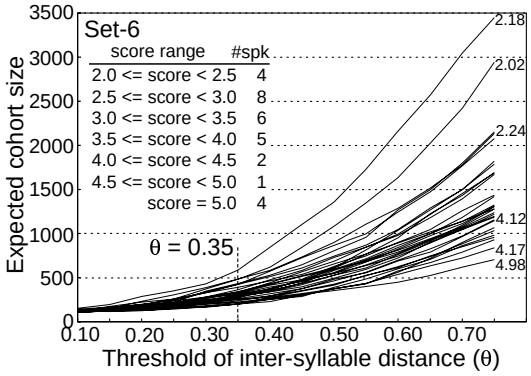

図 3: シラブル間距離 θ と ECS との関係

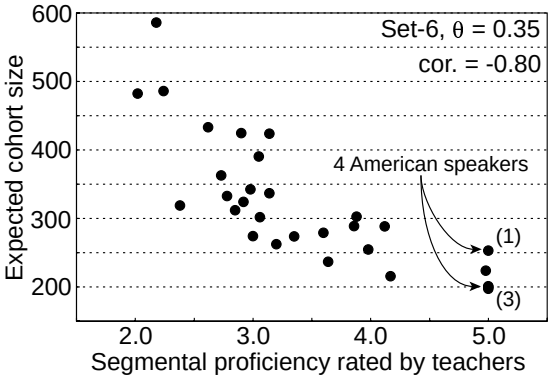

図 4: 自動発音評定結果

4.3 結果と考察

表 2 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，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 3 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させ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 3 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 4** 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，

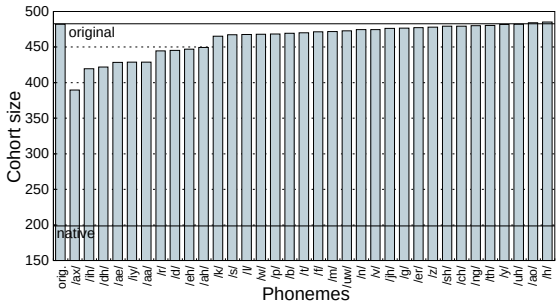

図 5: 一音素関係の置換による ECS の減少

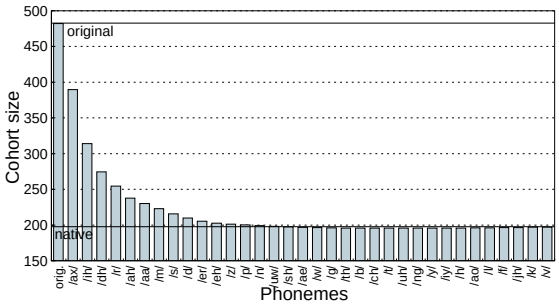

図 6: 逐次的な音素関係の置換による ECS の減少

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不一致問題が原理的に発生しない。

5　発音明瞭度向上のための最適な学習計画の自動生成

5.1　音響的普遍構造間における部分的構造の置換

音声の音響的普遍構造は，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そぎ落とすことが可能である (3)。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，学習者の構造に埋め込むことを考えると，これは，その音韻構造の獲得前・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・話者性などの情報までも置換することとなり，全く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。

5.2　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 $C = \{c_{ij}\}(1 \leq i, j \leq M)$ とする。「音素 p に関する関係」を $\{c_{pj}\}$ 及び $\{c_{ip}\}(1 \leq i, j \leq M)$ 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。音素数（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低減させる場合，次の学習対象を音素 p_0 と考える。

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 5** に示す。**第 4.2 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 6** 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。

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，収録環境など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。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。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。

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。

6　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，この音素構造と，対象言語の語彙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ることも示した。今後は更なる技術的安定性の追及，並びに，英語教師との議論を通して教材開発とその効果について検討する予定である。

参考文献

[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 (2003-3)

[2] 峯松信明他，“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，vol.27，no.3，pp.259–272 (2004-3)

[3] 峯松信明，松井健，広瀬啓吉，“音韻論の物理実装に基づく新しい音声の音響的表象”，電子情報通信学会音声研究会，SP2004-27，pp.47–52 (2004-6)

[4] 峯松信明，“外国語発音における音韻構造の歪みに着眼した発音の自動評定”，日本音声学会全国大会予稿集 (2004，掲載予定)

[5] Tsylb: http://www.nist.gov/speech/tools/tsylb2-11tarZ.htm