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

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

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，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。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り (1)，また，効率的なシステム開発を目的として，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）の構築も行なわれた (2)。

本研究では，発音能力の向上を支援する学習環境について検討する。従来開発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行なう形態となっている。しかし，現在の音声認識技術は「利用者・システム間の相性問題」を不可避的に伴う技術である。このような不可避的な不安定性は，技術の教育利用や医療利用において最も忌み嫌われる事項であることは想像に難くない。事実，最近になって，CALL システムの教育現場での使用について懐疑的な報告もされている (3)。

本研究では，近年峯松により提案されている音声の音響的普遍構造を用いて，学習者の発音状態を記述すること（即ち発音カルテの作成）により，上記不安定性の解決が可能であることを示す。音響的普遍構造は，構造音韻論の物理実装として位置付けられ，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」定義される新しい音声の物理表象である (4)。

2　音声に内在する音響的普遍構造

2.1　音声コミュニケーションに混入する不可避的な非言語情報のモデル化

音声認識・合成において，話者や音響機器特性に応じて音響モデルを適応させたり，音声特徴量を正規化する操作が広く行なわれている。この操作の基本事項を考慮することで，音声コミュニケーションに不可避的に混入する非言語情報をモデル化する。まず，非言語情報を三種類に分類する（加算性雑音，乗算性歪み，線形変換性歪み）。ラジオなどの背景雑音が加算性雑音の典型例であるが，本研究では，この種の雑音は考慮しない。物理的な抹消が可能であり，不可避ではないからである。

収録系，伝送系，再生系の音響機器による差異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ように，話者性の一部も乗算性歪みとなる。音声事象をケプストラム[1]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る（$c' = c + b$）。

話者が異なれば声道形状が，聴取者が異なれば聴覚特性が異なる。これらは線形変換性歪みの典型例である。声道形状・聴覚特性の差異はスペクトルに対する周波数ウォーピングとして近似されるが，(5) 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（$c' = Ac$）。

実際には様々な A や b が混入することになるが（A_i と b_i），最終的に得られる統合歪みも $c' = Ac + b$ として表現されることになる。即ち不可避な歪みは一次変換（アフィン変換）としてモデル化される。

2.2　情報理論に基づく音素群の構造化

構造音韻論では，個々の音素の属性を議論するのみならず，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，**図 1** に示す構造を提案している。以下，音素群を構造として見ることで非言語情報を表現する次元が数学的に消滅することを示す

[1]音声波形をフーリエ変換して得られる対数パワースペクトルを再度（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。

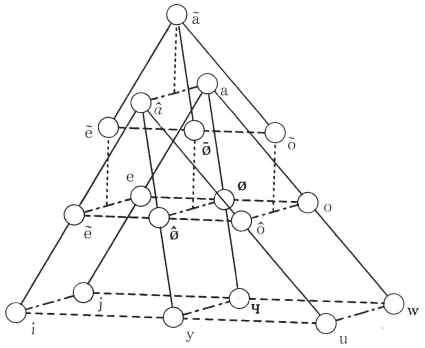

図 1: フランス語の音素群構造

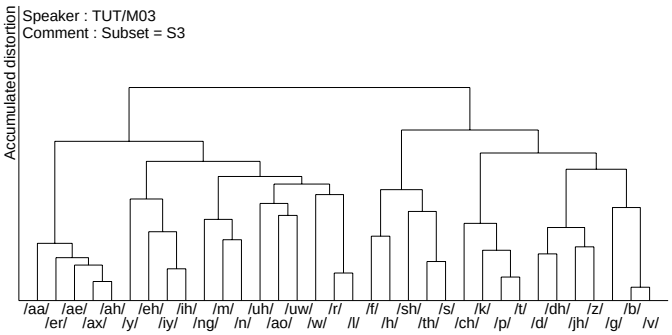

図 2: 日本人による英語音素樹型図

が，本研究では，素性を一切用いずに構造を考える。構造を規定するための幾何学的条件は，任意の二頂点間距離を全て求める（距離行列）ことであり，この距離行列を通して構造を捉える。

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，本研究では，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を考える。

2.3 構造として捉えられた音素群が持つ特徴

距離行列が構造と同値であること，不可避的な非言語情報がケプストラムの一次変換としてモデル化できることを述べた。音素群の構造化が非言語情報を表現する次元を消失させるためには，非言語情報の違いによって距離行列が不変であることを示せば良い。そして，それが（構造）音韻論研究者による（彼らの頭の中でも）非言語情報の抽象化の物理的・数学的実装となる。

前節で導入したバタチャリヤ距離は，二つの分布に対して同一の一次変換を施しても距離は不変であるという性質を持ち，距離行列の一次変換不変性が示される (4)。即ち不可避的に混入する非言語情報に依らず，普遍的に観測される物理的構造であり，これを音響的普遍構造と呼ぶ。音素を分布ではなく，音響空間内の点として定義した場合，一次変換によって一般に構造は歪んでくる。一次変換（アフィン変換）は画像（点の集合）を歪ませる変換として広く使われているが，分布の集合に対して，分布間距離を情報論的に定義すると，一切の歪みが消滅する。例えば声道長の長さによる差異（A）は構造の回転を意味し，マイクの違いによる差異（b）は構造の移動として観測される。なお，構造化することによる個人性の消失であるが，同一話者による異なる発声から得られる母音群構造差異と，異なる話者間による母音群構造差異とがほぼ同一であることが実験的に示されている (4)。

3 音素群構造に着眼した学習者発音の表象

音響的普遍構造に基づいて学習者発音を記述すれば，話者，年齢，性別，マイク，伝送系などの違いを示す次元が消失されるため，発音の是非のみを見定めることが可能である。と同時に，これを学習者別に行なえば，学習者の発音の変化の様子を記述する「文字」となる（発音カルテ）。**図 2** に音素群構造を樹型図として示した例を示す。ある学習によって，学習者がどのように変化するのか，をデータベースとして蓄積すれば，新しい学習者に対して効率的な学習過程を示すことも可能となる。また次節に示すように，二つの音素群構造間の距離（即ち学習者間の距離）をスカラー量として定義

可能である。即ち学習者行列が作成可能であり，これを用いれば学習者を分類する樹型図が描けることになる。地球上には 20 億人の英語学習者がいると言われるが，彼らを個人を単位として記述し[2]，全英語学習者を分類することが技術的には可能であると考えている。

4　音素群構造の比較に基づく外国語発音の自動評定

4.1　音素群より構成される構造間のスカラー差異の導出

N 次元ユークリッド空間内に M 個の点があるとする（$\{P_i\}, 1 \leq i \leq M$）。これらに対し M 点で張られる構造を考えることができるが，その重心を P_G とした場合，次式が真となる。

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

提案する音声表象では二つの分布間距離はバタチャリヤ距離で定義される。上記の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）群構造において，**式 (2)** の両辺は**図 3** に示す関係を呈する。以降，分布間距離としてバタチャリヤ距離の平方根を使用する。

次に二つ（二人）の構造 $\{P_i\}$，$\{Q_i\}$ を考える。但し同一の i に対応する音素は両者で等しいとする。様々な構造歪みが予測される日本人英語の母音，全音素構造に対し，下式が近似的に成立する。

$$\sqrt{\frac{1}{M^2}\sum_{i<j}\overline{P_iP_j}\times\overline{Q_iQ_j}} \approx \sqrt{\frac{1}{M}\sum_{i}\overline{P_iP_G}\times\overline{Q_iQ_G}} \tag{3}$$

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 4**（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 (2)，**式 (3)** より，以下の関係式が導かれる。

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \approx \sqrt{\frac{1}{M}\sum_{i}(\overline{P_iP_G}-\overline{Q_iQ_G})^2} \tag{4}$$

式 (4) 左辺は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 (4)** 右辺の物理的解釈を考える。今，二つの構造の重心が重なるように両者をシフトさせ（**図 5** 参照），その後，対応する点と点が重なるよう，一方の構造を重心周りに回転させて $\sum_i|\theta_i|$ を最小化する。なお，θ_i は**図 5** にあるように $\angle P_iGQ_i$ である。$|\theta_i|$ が十分に小さい場合は，

$$\left|\overline{P_iG}-\overline{Q_iG}\right| \approx \overline{P_iQ_i} \quad \text{（但し，シフト\&回転後）} \tag{5}$$

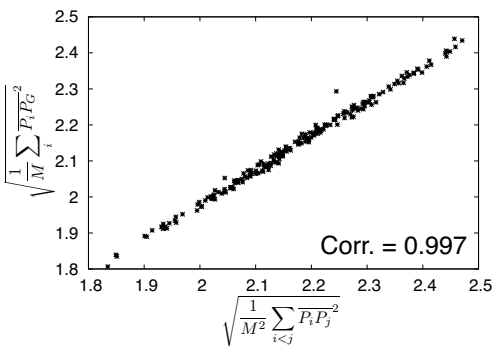

図 3: バタチャリヤ距離の平方根が持つ特性 1

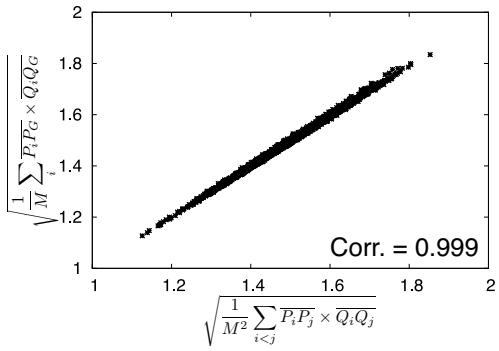

図 4: バタチャリヤ距離の平方根が持つ特性 2

[2]音声学は，音声活動を「音」を単位として記述し，音韻論は音声活動を「言語」を単位として記述する。その意味において，「個人」を単位として音声活動を描く方法論は，第三の音声科学と考えることも可能であろう。

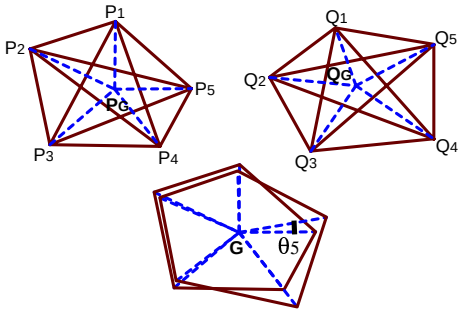

図 5: 構造の比較に基づく音響的照合

表 1: HMM 学習条件

サンプリング	16bit / 16kHz
窓	窓長 25 msc，シフト長 10 ms
パラメータ	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（PRONLEX 表記）

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

$$\sqrt{\frac{1}{M^2}\sum_{i<j}(\overline{P_iP_j}-\overline{Q_iQ_j})^2} \approx \sqrt{\frac{1}{M}\sum_i \overline{P_iQ_i}^2} \quad \text{（但し，シフト\&回転後）} \tag{6}$$

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 (6)** 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，音響モデルの適応を明示的に行なうことなく，構造間比較によって導かれることを意味する。

4.2 日本人・米国人に対する 2 話者間の構造歪み

式 (6) で定義された音素間距離行列を用いた 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 の各話者（米語話者含む）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。話者によって読み上げた文は異なっており，音素構造は読み上げた文サブセットに依存するため，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。

図 6 左図に母音群構造を用いた結果を示す。横軸は二話者間の構造歪みであり，縦軸は二話者間距離を HMM 間の距離の平均として求めたものである（つまり，音響空間内の絶対位置に基づく距離）。構造歪みでは，米国人・日本人間（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。**図 6** 右図に全音素構造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 間距離が接近している。これは 1) 本来 HMM 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

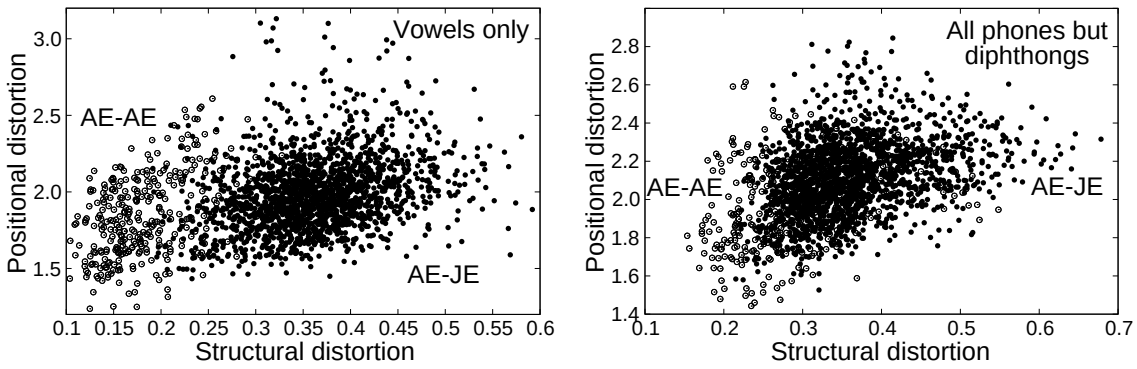

図 6: 構造に基づく歪みと絶対位置に基づく歪み（AE : American English，JE : Japanese English）

るが，ここでは対角化行列を使用している，3) モデルの近似誤差，などの実験条件の不具合も否めないが，ここでは，**式 (6)** の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。

4.3 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 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 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（教師・学生間の構造歪みを計算する場合，読み上げ文を揃える必要がある），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。**図 7** 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。良好な結果が得られている。

4.4 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 4.2 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 8** 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平均構造が得られる。今後，この構造分類の是非について，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，発話速度などの発話スタイル，母国語の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。

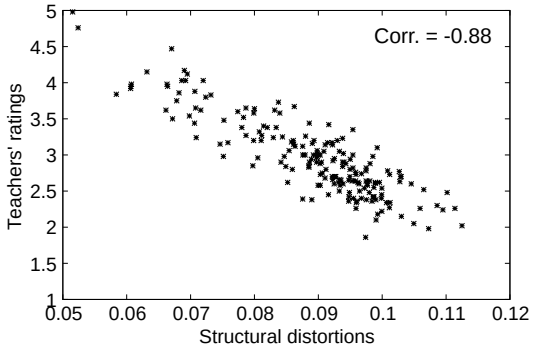

図 7: 構造歪みに基づく自動発音評定結果

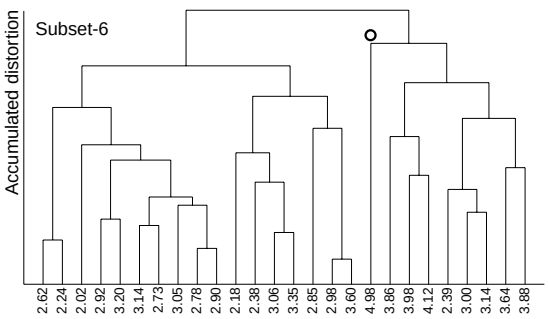

図 8: 構造歪みに基づく日本人英語の分類

5 まとめ

本研究では，発音評定技術の技術的安定性向上を目的として,「音声に不可避的に混入する非言語情報を表現する次元を保有しない」音声の物理表象として提案された音響的普遍構造を，外国語学習者の発音状態を表象する「文字」として利用することを提案した。そして構造として表象された学習者間（あるいは学習者と教師間）の差異をスカラー量として導出できることを示し，それを用いることで，発音の自動評定や，学習者群を分類することが可能であることを実験的に示した。本手法を用いることで，非言語情報がどの程度消滅するのか，と同時に，従来の発音評定技術と比較して，その安定性がどの程度向上したのかについては紙面の都合で示すことができなかったが，文献 (6) で詳細に述べているので，そちらを参考にして欲しい。また本手法は，あくまでも評定値を自動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 (6) では，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。

音声学は，世界中の言語に幾つの音が存在するのか，を調音的及び音響的に追及する音声科学であり，研究対象となる「音」の数は約 300 ほどである。一方音韻論は，世界中にどれだけの言語があり，各言語はどのようなシステムとして構成されているのか，を追及する音声科学であり，研究対象となる「言語」の数は約 3,000 ほどである。一方，外国語学習が本当に必要とする音声科学は,「音」や「言語」を単位として音声活動を観測する方法論ではなく,「個人」を単位とし，発音現象（本来は韻律的側面も含まれるべき）全体を網羅した形で「個人」を描く音声科学であると筆者は考えている[3]。そして地球上の全ての「個人」を対象とした場合，その数は 60 億となり，日々の学習・訓練によってその「個人」がどのように変化・改善されていくのかを逐一記述できる枠組みでなければならない。**第 3 節**でも述べたが,「個人」を対象とした音声科学を，第三の音声科学として位置づけることも可能であろう。そして本研究では，60 億の対象に対して（非言語情報をそぎ落とす形で）個々を（そしてその変化を）記述し，他者と比較し，全体を分類するための枠組みについて検討した。

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] A. Neri, *et al.*, “Automatic speech recognition for second language learning: How and why it actually works,” Proc. ICPhS, pp.1157–1160 (2003-8)

[4] 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)

[5] M. Pitz *et al.*, “Vocal tract normalization as linear transformation of MFCC,” Proc. EUROSPEECH, pp.1445–1448 (2003-9)

[6] 峯松信明，“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”，日本音声学会全国大会予稿集 (2004，掲載予定)

[3] ここで言う「個人」とは，性別，話者性，年齢といった「個人」を特定するための属性ではない。学習言語の発音に見られる母国語依存性は学習者一人一人異なるが，その依存性を指す。この場合上記属性は雑音以外の何者でもない。