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.

構造的特徴量に対する多段階の重回帰分析による発音評価

鈴木　雅之† 　中村　綾乃† 　喬　　宇† 　峯松　信明† 　広瀬　啓吉†

† 東京大学 〒 113–8656 東京都文京区本郷 7–3–1
E-mail: †{suzuki,ayano,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　国際化・情報化が急速に進む現在，コンピュータを用いた語学学習支援が盛んに研究され，教育現場に導入されている．語学学習のうち発音学習の支援には，子供から大人まで様々な声質を持つユーザにも頑健に動作する，精度の高い発音分析法が必要になる．しかし，現在広く用いられている Hidden Markov Model (HMM) を利用する発音分析法では，分析結果がユーザの声質と HMM の学習データの声質のミスマッチに影響されて，分析精度の低下をもたらす．近年この問題を解決するために，声質の違いに近似的に不変な音声の構造的表象を利用した発音分析法が提案され，ミスマッチに非常に高い頑健性を持つ発音分析が実現された．しかし，特にミスマッチが小さい場合，HMM を用いる手法に劣っている問題があった．本稿では，音声の構造的表象を用いた発音分析の精度を改善するために，多段階の重回帰分析を提案する．これにより，声質のミスマッチに高い頑健性を保ったまま，ミスマッチの小さい条件で HMM を利用する場合と同等以上の精度が得られた．さらに，提案手法と HMM を用いる手法を適切に組み合わせることで，さらに良好な結果が得られた．
キーワード　音声の構造的表象，CALL，重回帰分析，GOP

Pronunciation assessment based on multilayer multiple regression analysis using structural features

Masayuki SUZUKI†, Ayano NAKAMURA†, Yu QIAO†, Nobuaki MINEMATSU†, and Keikichi HIROSE†

† The University of Tokyo, 7–3–1, Hongo, Bunkyo-ku, Tokyo, 113–8656 Japan
E-mail: †{suzuki,ayano,qiao,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　In the rapid internationalization and informatization, many research efforts have been made to build computer-aided language learning (CALL) systems. Good pronunciation assessment systems should be built using the technologies which can deal with acoustic variabilities found in learners' utterances caused by non-linguistic factors such as age and gender. However, the widely-used acoustic modeling technique of HMM often shows unstable performances with speakers of different ages and genders. Recently, a new method of representing learners' pronunciations with their non-linguistic features effectively removed, called pronunciation structure. In this method, only the contrastive features of speech are extracted. However, the excessively high dimensionality of the structure comes to degrade its performance and, to solve this problem, multilayer regression analysis with structural features is proposed in this paper. The results show much higher correlation between human and machine performances of assessing learners' pronunciations compared to the previously proposed structure-based method. Further, the proposed method shows much higher robustness compared to the widely-used HMM-based method. In this paper, we also propose a good combination of the structure and the HMM.
Key words　speech structure, CALL, regression, GOP

1. はじめに

近年コンピュータを用いた語学学習（Computer Aided Language Learning; CALL）システムが広く用いられるようになった．この背景には，端末の普及により CALL システムが手軽に利用できるようになったことに加え，外国語教育への期待の高まりがある．例えば日本では，来年度より小学 5・6 年生の外国語活動が必修化される．ここで，小学校での英語活動は「話し言葉」としての英語であり，「話す／聞く」を中心に教育が展開される．このような状況から，今後「話す／聞く」教育をサポートする CALL システムのニーズは今後ますます高まっていくと考えられる．

しかし，子供から大人までどのような話者にも頑健に動作する発音分析システムの構築には，解決すべき課題が多い [1]．音声は，様々な声道形状を持つ話者に，様々な環境下で発声され，様々な録音機器により収録される．これらのプロセスの中で，音声の物理的実体は，たとえ同じ発音の情報を持っていたとしても，非言語的特徴によって変形されてしまう．そのため，あるユーザの発音とある母語話者の発音を比較してどこに違いがあるか分析しようとしても，非言語的特徴のミスマッチにより，発音の間違いのみを正しく見つけることが困難になってしまう．

従来の発音分析技術は，大量の音声データから，隠れマルコフモデル（Hidden Markov Model; HMM）による音響モデルを学習し，それを話者正規化・適応させることでこの「ミスマッチ問題」を解決しようとしてきた [2]．しかしこの方法では，話者適応用のデータそのものに発音誤りが含まれるため，生徒の声質のみならず発音誤りにも適応がかかってしまい，誤った発音を正しいと判断してしまう問題が起こりうる [3]．この問題は，非言語的特徴の違いと発音の情報の違いがまったく区別されていないために起こる問題である．

近年，音声に含まれる発音の情報のみを表現するために，静的な非言語的特徴に理論的に不変な音声の構造的表象が提案された [4]．音声の構造的表象を用いることで，非言語的特徴のミスマッチに頑健な音声アプリケーションを実現することが可能となる．既に，発音分析・評価 [5]，孤立単語音声認識 [6] 等への応用が研究され，そのミスマッチに対する頑健性が実験的に示されている．しかしながら，音声の構造的表象を用いた発音分析には，精度に問題があった．特に非言語的特徴のミスマッチが比較的小さい場合には，従来の HMM を用いる手法と比較して精度が劣っていた．

そこで本稿では，音声の構造的表象を用いた発音分析の精度を向上させるために，多段階の重回帰分析を提案する．提案手法により，ミスマッチに対する高い頑健性を保ったまま，ミスマッチが小さい場合にも HMM を用いた手法と同等以上の精度が得られた．さらに，構造的表象を用いる手法と HMM を用いる手法を適切に組み合わせることにより，さらに高い精度が得られることも示す．

2. 音声の構造的表象

2.1 音声に含まれる非言語的特徴

音声分析の特徴量として広く用いられているケプストラムには，発音の情報と非言語的特徴が同時に符号化されている．我々の目的は，ケプストラムから非言語的特徴を取り除き，発音の情報のみを抽出することである．そこでまず，非言語的な特徴がケプストラムをどのように変動させるのか考える．

多くの非言語的特徴は，ケプストラム c に対する静的な可逆変換 $c' = h(c)$ で近似することができる．例えば音声から言語情報を保持したまま話者性のみを変換する技術である話者変換の研究では，話者の違いをケプストラムの可逆な非線形変換と仮定し，その変換関数を GMM（Gaussian Mixture Model）を利用して学習する手法が広く用いられている [7]．また，音声認識における音響モデルの MLLR（Maximum Likelihood Linear Regression）適応では，話者の違いを線形変換と仮定した上でその変換パラメータを推定することで実現されている．以上のことから，話者の違いはケプストラムに対する静的な可逆変換 $c' = h(c)$ で近似できることがわかる．なお声質の他にも，録音機器や伝送経路の周波数特性の違いや，静的な背景雑音も $c' = h(c)$ で表現することができる．

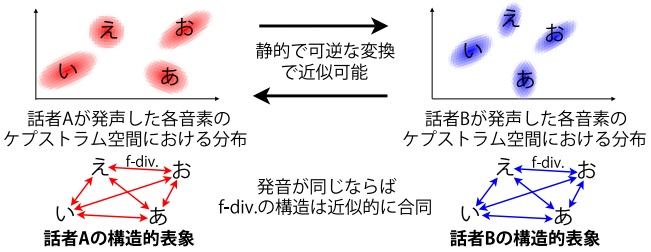

図 1　静的で可逆な変換に不変な音声の構造的表象

2.2 音声の構造的表象

音声に含まれる様々な非言語的特徴のうち，声質の違い，録音機器の違い，定常雑音の違いに関しては，静的な可逆変換 $c' = h(c)$ で表現できる．ここで，変換 h の前後で不変な特徴量が見つかれば，その特徴量はこれらの非言語的特徴に影響を受けないということができる．

二分布間距離尺度である f-divergence は，このような変換 h に不変となる [8]．二つの分布 p_i, p_j 間の f-divergence は以下の汎関数で表される．

$$f_{div}\left(p_i, p_j\right) = \oint p_j(x)\, g\left(\frac{p_i(x)}{p_j(x)}\right) dx \qquad (1)$$

ただし，$g(x)$ は $x > 0$ で定義される凸関数であり，$g(1) = 0$ を満たすとする．可逆な変換 $c' = h(c)$ により，$p_i(c), p_j(c)$ がそれぞれ $q_i(c'), q_j(c')$ に変換される場合の f-divergence の不変性は，$J(c')$ を $h^{-1}(c')$ のヤコビアンの行列式の絶対値として，以下のように証明できる．

$$f_{div}(p_i, p_j) = \oint p_j(c) g\left(\frac{p_i(c)}{p_j(c)}\right) dc \qquad (2)$$

$$= \oint p_j(h^{-1}(c')) g\left(\frac{p_i(h^{-1}(c'))J(c')}{p_j(h^{-1}(c'))J(c')}\right) J(c')dc' \qquad (3)$$

$$= \oint q_j(c') g\left(\frac{q_i(c')}{q_j(c')}\right) dc' = f_{div}(q_i, q_j) \quad \square \qquad (4)$$

以上により，f-divergence が変換 h に不変になることが証明された．なお証明は省くが，可逆な変換 $c' = h(c)$ に不変な分布間距離尺度は f-divergence でなければならないという必要性も証明することができる [8]．

図 1 のように，ケプストラム空間において，音素などの音響イベントを分布化し，それらの間の f-divergence を計算することで一つの構造が得られる．これを，音声の構造的表象と呼ぶ．音声の構造的表象を構成する各エッジは f-divergence であるので，音声の構造的表象は可逆変換 $c' = h(c)$ で表現できる非言語的特徴に理論的に不変となる．

具体的に f-divergence を計算するためには，適当な $g(x)$ を

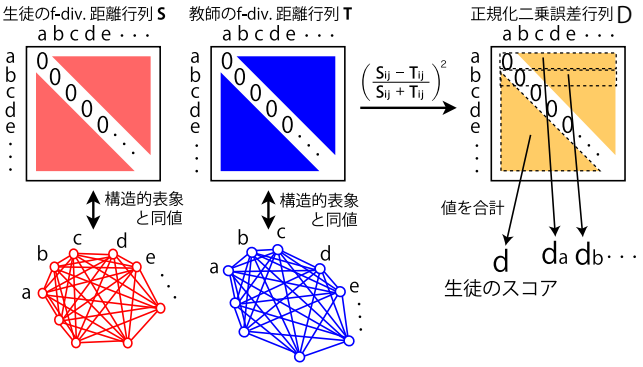

図 2　従来の構造的表象を用いた発音分析

決定する必要がある．本研究では，$g(x) = \sqrt{x}$ とおき，その $-\ln$ をとった距離尺度であるバタチャリヤ距離の平方根を構造的表象の抽出に利用する．

2.3　従来の構造的表象を用いた発音分析

音声の構造的表象を用いて外国語発音分析を行うためには，ユーザの発声から抽出した音声の構造的表象と教師の発声から抽出した音声の構造的表象を比較し，どの程度差異があるかを見ればよい．ここで，二つの構造間差異 d は，生徒の f-divergence 距離行列を S，教師の f-divergence 距離行列を T として，

$$d = \sum_{i<j} D_{ij} = \sum_{i<j} \left(\frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right)^2 \qquad (5)$$

が利用されている [9]．生徒と教師の構造的表象から正規化二乗誤差行列（D，図 2 参照）を計算し，(6) により d を計算することで発音評価を行う枠組みを図 2 に示す．なお正規化二乗誤差行列は，上三角部分の和をとって生徒のスコアとするだけでなく，要素ごとの値を見ることで，音響イベントごとの評価も可能である．例えば，音響イベント a に関する構造間差異 d_a として，下式で計算される正規化二乗誤差行列の各行の和が利用できる．

$$d_a = \sum_{i} D_{ai} = \sum_{i} \left(\frac{S_{ai} - T_{ai}}{S_{ai} + T_{ai}} \right)^2 \qquad (6)$$

構造的表象は，あらゆる静的で可逆な変換に対して不変である．そのため構造的表象を用いた発音分析は，CMN（Cepstrum Mean Normalization），最適な SS（Spectral Subtraction），最適な VTLN（Vocal Tract Length Normalization），最適な MLLR 適応などを内包する効果を，それらを行うことなしで得ることができる．

3.　構造的特徴と多段階の重回帰分析

3.1　二段階重回帰分析

音声の構造的表象を表現する特徴の次元数，すなわち構造のエッジの数は，音響イベントの数を M として ${}_MC_2$ である．そのため，子音母音を両方含む発音分析では M が大きくなり，二乗オーダーで構造の次元数が大きくなってしまう．具体的に米語発音を分析することを考えた場合，米語の音素は 42 個存在するため，音響イベントを音素とすると構造的特徴の次元は

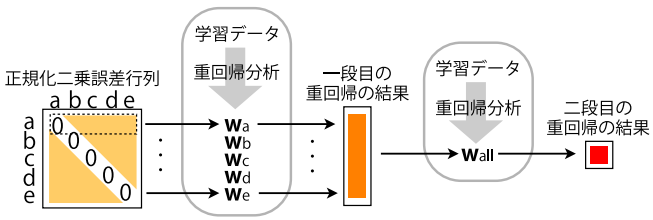

図 3　構造的特徴と二段階重回帰分析を用いた発音評価

${}_{42}C_2 = 861$ 次元となってしまう．次元数が高くなりすぎると，「次元の呪い」の問題により，分析精度が低下してしまう．

このような場合，次元圧縮を用いることが有効であると考えられる．しかし，単純に主成分分析（Principal Component Analysis; PCA）や判別分析（Linear Discriminant Analysis; LDA）などを用いた次元圧縮を用いるには問題がある．スパースでない 861 次元の特徴量を次元圧縮するためには，最低でも 861 個のデータが必要になり，十分な精度で次元圧縮を行うにはさらに多くのデータが必要となる．データが十分に得られない場合，過学習がおき，次元圧縮の精度は十分でなくなってしまう．さらに，PCA や LDA では，次元圧縮後の特徴量の解釈ができなくなってしまう問題もある．音響イベントごとの評価として利用できる正規化二乗誤差行列の各行の和など，行列の配置そのものが発音分析に有効な情報であるのに，PCA や LDA を用いると，その情報が消えてしまう．

そこで，少ないデータで学習できてかつ次元圧縮後の意味解釈の行いやすい次元圧縮手法として，二段階重回帰分析を提案する．構造的特徴と二段階重回帰分析を用いた外国語発音評価の枠組みを，図 3 に示す．ここで，w_i は重回帰分析により推定された重みベクトルである．二段階重回帰では，正規化二乗誤差行列を二段階にわけて重回帰分析を行う．二段階で重回帰分析を行うことにより，学習するパラメータを大幅に減らして過学習を防ぎ，かつ途中段階で音響イベントごとの評価値を残しながら次元圧縮を行うことができる．

1 段目の重回帰分析では，正規化二乗誤差行列の行ごとに重回帰分析を行う．重回帰分析の目的変数には，各音素に対する手動評価値を利用する．このように重回帰分析の重みパラメータを学習すると，ある音響イベントの発音を評価する際にどの音響イベントとの相対関係を重要視するかが学習される．例えば，日本人が発音する米語の/s/を評価する際には，日本人が/s/と混同しやすい/sh/や/th/などとの相対関係が重要視されると予想される．1 段目の重回帰の結果は，各音響イベントごとの評価値として利用できる．

2 段目の重回帰分析では，1 段目の重回帰で得られる各音響イベントごとの値を説明変数にして重回帰分析を行う．重回帰分析の目的変数には，生徒に対する手動評価値を利用する．このように重回帰分析の重みパラメータを学習すると，生徒の発音全体のスコアを算出する際に，どの音響イベントを重要視するかが学習される．例えば，日本人の米語発音を評価する場合には，日本人が一般的に苦手にしやすい/r/や/er/などが重要視されると予想される．2 段目の重回帰の結果，生徒の評価値

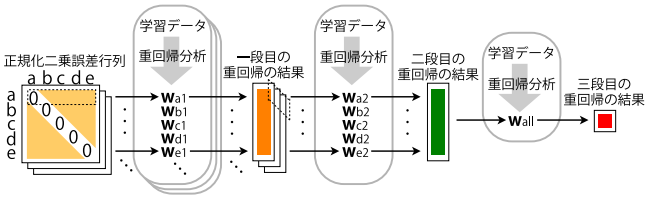

図 4　構造的特徴と三段階重回帰分析を用いた発音評価

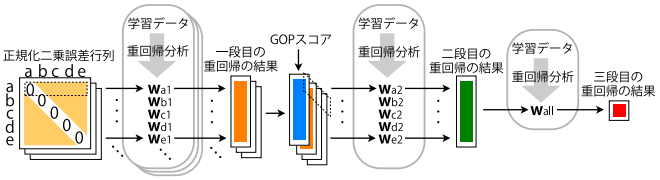

図 5　構造的特徴と GOP スコアを用いた発音評価

が得られる．

3.2　複数の構造的表象と三段階重回帰分析

ここまでは，一つの正規化二乗誤差行列を使って発音分析を行っていた．そのため，比較している教師の方言や発音の癖なども含めて，生徒の発音とどの部分が異なるのかを分析することになる．ここで，教師を複数用意して，生徒と複数の教師の構造的表象を一対他で比較すれば，複数の正規化二乗誤差行列が得られる．複数の正規化二乗誤差行列を使って発音分析を行うことで，教師の方言の違いや発音の癖をある程度吸収することが可能になる．また，教師を複数用いる以外にも，音響特徴量を変更すれば複数の正規化二乗誤差行列が得られる．例えば，音声分析には多くの場合 16kHz サンプリングの音声が用いられるが，母音に関しては母音を特徴付けるフォルマント周波数はおよそ 3kHz 以下に存在しており，逆に高い周波数成分には話者の違いが強く現れるといった報告がある [10]．ここで，通常の 16kHz サンプリングの音声と，3kHz ローパスフィルタをかけた音声を用意し，それぞれから構造的表象を抽出すれば，二つの正規化二乗誤差行列を得ることができる．これらを使って発音分析することで，特に母音に関しての精度向上が見込める．以上のように，教師や音響特徴量を増やすことで，正規化二乗誤差行列を複数得て，情報量を増やすことができる．ただし情報量が増えると同時に，次元数は正規化二乗誤差行列の枚数を N として N 倍になり，次元の呪いの問題がさらに強くなってしまう欠点もある．

そこで，二段階重回帰分析を拡張した，三段階重回帰分析により，複数の正規化二乗誤差行列から次元の呪いを適切に避け，情報量が増える利点を生かす手法を提案する．三段階重回帰分析の枠組みを図 4 に示す．1 段目の重回帰分析と 3 段目の重回帰分析は，それぞれ二段階重回帰分析の 1 段目と 2 段目に相当する重回帰分析を行う．

2 段目の重回帰分析では，1 段目の重回帰で得られる各音響イベントごとの複数のスコアを説明変数にして重回帰分析を行う．重回帰分析の目的変数には，1 段目の重回帰分析と同じく，各音響イベントごとの手動評価値を利用する．このように重回帰分析の重みパラメータを学習すると，ある音響イベントの発音を評価する際に，どの音響特徴量，もしくはどの教師との差異を重要視するかが学習される．例えば，16kHz サンプリング音声と 3kHz ローパスフィルタをかけた音声を用いる場合，3kHz 以下に音素の特徴が多く含まれる母音はローパスフィルタをかけた構造が重要視され，高周波数領域にも音素の特徴が多く含まれる子音は 16kHz サンプリングの構造が重要視されると予想される．2 段目の重回帰の結果は，各音響イベントごとの評価値として利用できる．

3.3　HMM を用いる手法との組み合わせ

音声の構造的表象は，音の絶対的な特徴を捨て，音と音との相対関係のみを記述している．一方 HMM を用いる手法では，音そのものの絶対的な特徴を記述している．そのため，両者は音声の異なる特徴を捉えており，両者を組み合わせることでさらに精度の高い分析が行えると考えられる．例えば，母音のような話者の影響を強く受ける音素は，構造的表象を用いて話者不変な相対的な特徴を見た方がよいと考えられる．一方，無声子音のような話者によって大きな変化がない音素では，HMM を用いて音そのものの絶対的な特徴を利用した方がよいと考えられる．

そこで，構造的表象と HMM を用いる手法を適切に組み合わせて利用する手法を提案する．図 5 に，提案手法を示す．なお，HMM を用いた発音評価手法の一つとして，GOP（Goodness Of Pronunciation）スコアを採用する [11]．GOP スコアは，以下の近似式を用いて計算される．

$$
\begin{aligned}
&GOP(o_1,...,o_T,p_1,...,p_N)\\
&\quad=\log P(p_1,...,p_N|o_1,...,o_T) &(7)\\
&\quad\approx\frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\sum_{q\in Q}P(o^{p_i}|q)}\right\} &(8)\\
&\quad\approx\frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\max_{q\in Q}P(o^{p_i}|q)}\right\} &(9)
\end{aligned}
$$

ここで，T は観測フレーム長であり，N は意図された音素数である．o^{p_i} は強制アライメントによって得られた音素 p_i に対応する音声区間であり，D_{p_i} はその区間長である．$\{o^{p_1},\cdots,o^{p_N}\}$ と $\{o_1,\cdots,o_N\}$ は同一の音響特徴量時系列を表現しており，Q は全音素集合である．(9) の Σ の中の式を計算すれば，各音素ごとの GOP スコアを得ることができる．

GOP スコアを三段階重回帰分析の 2 段目の重回帰分析の説明変数に加えることで，無声子音など音そのものの絶対的な特性が有効な音素では GOP スコアに対する重みパラメータが大きくなり，逆に母音などでは構造的表象に対する重みパラメータが大きくなると考えられる．

4.　実　　験

4.1　日本人英語読み上げ音声データベースを用いた実験

提案手法の有効性を検証するため，日本人英語読み上げ音声（English Read by Japanese; ERJ）データベースを用いた実

表 1　音響分析条件

サンプリング	16bit / 16kHz
窓	25 msec 幅，10 msec シフトのハミング窓
学習データ	一名につき 60 文（set 8 の話者のみ 40 文）
音響特徴量	MFCC(12)+Energy+ΔMFCC(12)+ΔEnergy
HMM	特定話者音素 monophone HMM
出力確率分布	対角共分散行列を持つガウス分布
トポロジー	3 状態の left to right 型
音素の種類	ɑː,æ,ʌ,ɔː,aʊ,ə,ər,aɪ,b,ʧ,d,ð,ɛ,ɚː,eɪ,f,g,h,ɪ,iː,ʤ,k,l,m, n,ŋ,oʊ,ɔɪ,p,r,s,ʃ,t,θ,ʊ,uː,v,w,y,z,ʒ,sil 合計 42 種類

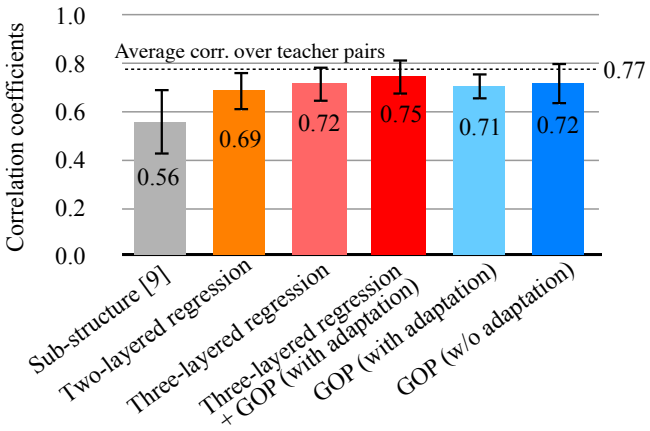

図 6　手動評価値との相関の平均値と標準偏差

験を行う [12]．ERJ データベースは，200 名の日本人大学生と米語母語話者 20 名が英語文章を読み上げた音声が収録されている．読み上げられた文章セットには，8 つの種類があり，各セットには 40 文もしくは 60 文が含まれている．母語話者は，文セットのうち 4 セット分を読み上げている．ただし，20 名中 2 名（男性 M08 と女性 F12）は，全 8 セット分を読み上げている．日本人大学生は，1 セット分のみを読み上げている．さらに，各学生が読み上げた文章のうち，1 人につき 10 文に対して発音の手動評価値が付与されている．手動評価者は日本人学習者の癖をよく理解している，米語母語話者の音声学者 5 名で，音素的に正しく発声されたかについて，5 段階で評価されている．

構造的表象を抽出するための音響分析条件を表 1 に示す．まず，生徒の読み上げ音声それぞれから，米語音素 42 個分の特定話者音素 HMM を作成し，HMM の出力確率分布の間の f-divergence 距離行列を計算することで構造を抽出する．HMM のトポロジーとして 3 状態の left-to-right 型 HMM を用いているため，音素を 3 分割したものが一つの分布になっているが，二つの音素 HMM の対応する 3 つの分布間の f-divergence の和を一つのエッジとして構造的表象を作成する．これにより，音素が音響イベントになり，${}_{42}C_2 = 861$ 本のエッジからならなる構造的表象が得られる．次に，同様の読み上げ文セットを学習データ，同様の条件で教師の構造も抽出する．ここで教師の音声には，男性教師 M08 1 名のみを利用した．

このように計算した構造的特徴を用いて提案手法である多段階重回帰分析を用いた発音評価を行う．重回帰分析の学習データには，8 セットのうち 7 セット分の音声を用い，8-fold cross validation で評価を行った．重回帰分析には，二次の正則化項を導入した重回帰分析であるリッジ回帰分析を用いた．リッジ回帰にもちいる正則化パラメータ λ は，予備実験よりすべて 1 と設定した．重回帰分析の目的変数には，ERJ の各話者につけられた 10 文 ×5 名の手動評価値の平均を利用した．また音素ごとの手動評価値は ERJ データベースに付属していないため，多段階重回帰の全段階の重回帰分析の目的変数も同様に評価値の平均を利用した．また三段階重回帰分析におけるマルチストリーム化の際には，教師の音声に M08 と F12 の二名を，さらに表 1 の条件の MFCC の他 3kHz ローパスフィルタを通した音声の MFCC と二種類の音響特徴量を用い，2×2 の 4 ストリームの正規化二乗誤差行列を用いた．また GOP スコアを算出するための HMM には，ERJ に含まれるネイティブ話者 20 名分の音声を用いて学習し，MLLR 適応をかけて利用した．さらに比較実験として，[9] で提案している部分構造を用いる実験と，GOP スコアのみを用い音素ごとの GOP スコアを多段階重回帰分析と同様の条件で重回帰分析した実験も行った．

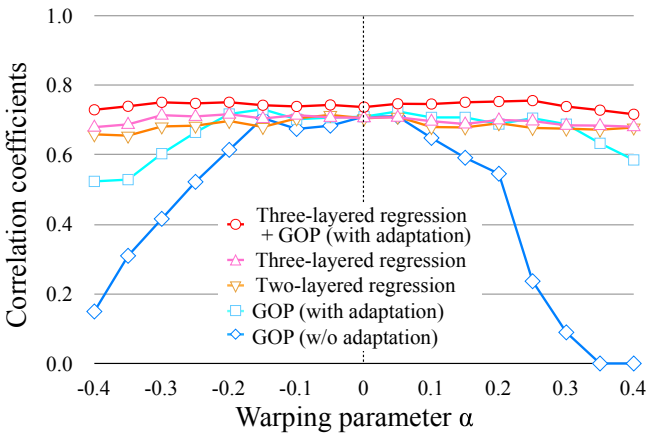

図 7　声道長変換をかけた音声の手動評価値との相関の平均値

結果を図 6 に示す．参考に，5 名の音声学者の手動評価値間の相関の平均値も示す．結果，多段階重回帰分析により大幅に相関が向上していることがわかる．また，二段階重回帰分析よりも三段階重回帰を用いた方がわずかに相関が向上し，さらに GOP スコアと組み合わせることでさらに相関が向上していることがわかる．また，GOP スコアを重回帰分析してスコアと比較しても，提案手法は同程度以上の相関になっている．

4.2　ミスマッチ条件下での実験

提案手法のミスマッチに対する頑健性を調べるため，ERJ データベースに含まれる生徒の読み上げ音声に，人工的な声道長変換をかけた音声の分析実験を行った．声道長変換には，声道長を変化させる周波数ウォーピング関数として 1 次の全域通過フィルタ

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{10}$$

による近似 [13] を利用し，これを音声分析変換合成法である STRAIGHT を利用して実装した [14]．このとき声道長の変換度合いは，ウォーピングパラメータ α によって調整される．た

表 2　実 験 結 果

ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː	総合
0.57	0.55	0.74	0.03	0.04	0.73	0.82	0.86	0.73	0.55	0.88

表 3　手動評価値間の相関の平均値

ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː	総合
0.71	0.68	0.52	0.47	0.16	0.82	0.66	0.68	0.47	0.48	0.52

だし，$|\alpha| < 1$ であり，$\alpha = 0$ のときに変換なし，$\alpha = \pm 0.40$ の時に，声道長が 1/2・2 倍になることにおよそ対応している．

先の実験と同じ条件を用い，提案手法である二段階重回帰分析，三段階重回帰分析，三段階重回帰分析+GOP スコアを用いた実験と，比較実験として GOP スコア単独を用いた実験，さらに参考のために MLLR 適応をかけていない HMM を用いて算出した GOP スコアを用いた実験を行った結果を図 7 に示す(注1)．まず提案手法を用いた場合には，声道長を変化させてもほとんど相関は変化していない．すなわち，ミスマッチに非常に頑健な処理が行えていることが分かる．一方，MLLR 適応を用いていない GOP スコアは，声道長の変化により相関が大幅低下してしまっている．MLLR 適応を用いると，ある程度頑健性がでてくるが，それでも大きく声道長を変化させた場合には相関が若干低下している．この結果から，提案手法は MLLR 適応より声道長変換に対して強い頑健性を持つことがわかる．これは，構造的表象を用いる手法が，最適な MLLR 適応を内包する効果を持っていることに対応している．

4.3　音素ごとのスコアの推定実験

最後に，多段階重回帰分析の途中段階で得られる音素ごとのスコアの精度を検証する実験を行う．音声データとして，日本人中学生 36 名と 5 名の英語教師が発声した，米語単母音を網羅する 10 単語セット（pot[ɑː], bat[æ], but[ʌ], bought[ɔː], bet[ɛ], bird[ɚː], bit[ɪ], beat[iː], put[ʊ], boot[uː]）の読み上げ音声を収録した．この音声データに対し，音声学者 4 名が，各母音それぞれのスコアおよび総合スコアを，4 段階評価している．この単語音声データから，HMM の強制アライメントにより母音部分のみを切り出し，その平均と分散を計算することで各母音を分布化し，構造的特徴を抽出した [5]．これを用いて多段階重回帰分析による発音評価を行い，手動評価値との相関値を調べた．

結果を表 2 に示す．結果，多くの母音において十分に高い相関が得られていることがわかる．参考に，音声学者間の手動評価値の相関の平均値を表 3 に示す．表 3 の相関がおよそ音声学者のパフォーマンスということができるが，提案手法は母音によっては音声学者と同程度の精度で発音分析を行えている．なお，表 2，表 3 で共に相関が低い母音（ɔː，ɛ，uːなど）は，日本語の母音がほぼそのまま代用できる母音であり，そもそもスコアの分散が低くなっていることが原因である．

5. おわりに

本論文では，構造的表象を用いた外国語発音評価において，適切な自由度と汎化性能をもち意味解釈の行いやすい次元圧縮法として，多段階重回帰を提案した．実験の結果，従来の構造的表象用いた外国語発音評価手法と比較して，大幅に精度が向上させられることがわかった．また，GOP スコアを利用した手法と比較しても，提案手法はより高い精度が得られた．さらに GOP と提案手法を組み合わせることで，より高い精度が得られた．

(注1)：STRAIGHT の分析再合成をかけているため，$\alpha = 0$ で声道長変換をかけていない場合でも，先の実験と完全に同じ結果にはなっていない．

文　献

[1] M. Russell and S. D'Arcy. Challenges for computer recognition of children's speech. In *Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE'2007)*, 2007.

[2] Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, and S. Makino. A speaker adaptation method for non-native speech using learners' native utterances for computer-assisted language learning system. *Speech Communication*, Vol. 51, No. 10, pp. 875–882, 2009.

[3] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose. Analysis and utilization of mllr speaker adaptation technique for learners' pronunciation evaluation. In *Proc. INTERSPEECH'2009*, pp. 608–611, 2009.

[4] N. Minematsu. Yet another acoustic representation of speech sounds. In *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP'2004)*, pp. 585–588, 2004.

[5] 朝川智, 峯松信明, 広瀬啓吉. 音声の構造的表象に基づく英語学習話者発音の音響的分析. 電子情報通信学会論文誌, Vol. J90-D, No. 5, pp. 1249–1262, 2007.

[6] Y. Qiao, S. Asakawa, N. Minematsu, and K. Hirose. On invariant structural representation for speech recognition: theoretical validation and experimental improvement. In *Proc. INTERSPEECH*, pp. 3055–3058, 2009.

[7] 戸田智基, 陸金林, 猿渡洋, 鹿野清宏. 周波数軸伸縮を用いた混合正規分布モデルに基づく声質変換法. 電子情報通信学会論文誌, Vol. J84-D-II, No. 10, pp. 2181–2189, 2001.

[8] Y. Qiao and N. Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Trans. on Signal Processing*, Vol. 58, No. 7, pp. 3884–3890, 2010.

[9] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose. Substructure-based estimation of pronunciation proficiency and classification of learners. In *Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU'2009)*, pp. 574–579, 2009.

[10] T. Kitamura, K. Honda, and H. Takemoto. Individual variation of the hypopharyngeal cavities and its acoustic effects. *Acoustical Science and Technology*, Vol. 26, No. 1, pp. 16–26, 2005.

[11] S.M. Witt and S. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, Vol. 30, pp. 95–108, 2000.

[12] 峯松信明, 富山義弘, 吉本啓, 清水克正, 中川聖一, 壇辻正剛, 牧野正三. 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.

[13] 江森正, 篠田浩一. 音声認識のための高速最ゆう推定を用いた声道長正規化. 電子情報通信学会論文誌, Vol. J83-D-II, No. 11, pp. 2108–2117, 2000.

[14] H. Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. & Tech.*, Vol. 27, No. 6, pp. 349–353, 2006.