情報処理学会論文誌 Vol. 52 No. 5 1899–1909 (May 2011)

音声の構造的表象と多段階の重回帰を用いた外国語発音評価

鈴木雅之[†1] 峯松信明[†1] 広瀬啓吉[†1]

国際化・情報化社会の現在，コンピュータを用いた語学学習支援がさかんに行われている．語学学習のうち発音学習の支援には，子供から大人までどのような声質を持つ学習者を対象とした場合でも頑健に動作する，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，学習者の声質と HMM の学習データの声質のミスマッチに影響されて，発音分析としては不適切な結果を呈することがある．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．しかし頑健性は高いものの精度は不十分であり，特にミスマッチが小さい場合，HMM を用いる手法と比較して十分な精度が得られていなかった．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．

Non-native Pronunciation Assessment Based on Pronunciation Structure and Multilayer Regression

Masayuki Suzuki,[†1] Nobuaki Minematsu[†1] and Keikichi Hirose[†1]

In the rapid internationalization and informatization, many research efforts have been made to build Computer-Aided Language Learning (CALL) systems for language learners. What is requisite for good CALL systems is robust analysis of pronunciation with regard to acoustic changes caused by non-linguistic factors such as age and gender. However, the widely used acoustic model, which is based on HMMs, often shows an unstable performance with speakers of different age and gender. Recently, to solve this problem, a new method for modeling learners' pronunciations was proposed, called pronunciation structure. In this method, the non-linguistic features are effectively removed. The pronunciation structure shows a much more robust performance in mismatched conditions but its performance was lower than that of HMMs in matched conditions. To solve this problem, multilayer regression analysis is introduced to the pronunciation structure. The experimental results show a much higher performance compared to the previously proposed structure-based method, which is comparable to the HMM-based method in matched conditions. Further, by combining the structure-based method and the HMM-based method, we realize an even higher and more robust performance.

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．たとえば日本では，文部科学省が 2002 年度に『「英語が使える」日本人育成に向けた戦略構想』を策定し，2011 年度より小学 5・6 年生の外国語活動を必修化することを決定した．約 240 万人の児童が新たに英語を学習することになり，約 8 万人の新たな英語教師が求められている．しかし，英語教師の数的拡充は困難であり，文科省はクラス担任に英語教育の中核となるよう要請している[1)]．すなわち，英語を専門としない教師たちが英語の授業を担当する．なお，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」教育が実施される．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声質を持つ話者にも，その声質によらず頑健に動作する発音分析システムの構築には，大きな技術的問題がある[2)]．音声は，様々な声道形状を持つ話者によって様々な環境下で発声され収録される．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，ある学習者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，上記のような非言語的要因によるミスマッチにより，発音の間違いのみを正しく見つけることが困難となる．

従来の発音分析技術は，多くの話者の多様な環境下の音声データを集め，HMM（Hidden

†1 東京大学
The University of Tokyo

Markov Model）による音響モデルを学習し，それを少量の学習者の発声を用いて話者適応させることで「ミスマッチ問題」を解決しようとしてきた[3]．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，学習者の声質だけでなく発音誤りにも適応がかかってしまい，適応後の HMM を用いると誤った発音を正しいと判断してしまうという問題が生じうる[4]．この問題は，非言語的特徴の違い（たとえば話者の違い）も，個々の音韻の発音の違い（たとえば [r] と [l] の違い）も，物理的にはスペクトル包絡の違いとして観測されるが，この両者を区別していないために起こる問題である．

近年，発音の情報のみを表現するために，多くの非言語的特徴に近似的に不変な音声の構造的表象が提案された[5]．これは，音声の物理的・絶対的実体を捨て，それらの相対関係のみをとらえることによって得られるものである．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．すでに，発音分析・評価[6]，孤立単語音声認識[7] など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十分な精度が得られていない．たとえば構造的表象を用いた音声認識の場合，特に子音を扱うとき，母音のみを扱うときと比べ大きな精度劣化が観測されている[8),9)]．同様の問題は，発音分析でも生じうると考えられる．

そこで本稿では，音声の構造的表象を用いた，母音子音すべてを含む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．提案手法を用い，構造的表象を構成する各エッジに適切な重み係数を掛けることで，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，より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

音声分析の特徴量には，たとえば MFCC（Mel Frequency Cepstrum Coefficients）など，対数パワースペクトル領域の特徴量を逆周波数変換した，ケプストラム領域の特徴量が広く用いられている．ケプストラムには，発音の情報と，年齢・性別・話者性などの非言語的特徴が同時に含まれている．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について見る．

非言語的特徴の中でも，最も違いが大きくかつ本質的に不可避なものが，声質の違いである．声質の違いは，話者の性別や年齢などによる，声道形状の違いによって発生する．声質の違いは，時刻 t におけるケプストラム $\boldsymbol{c}_t$ に対する可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ で近似できることが知られている．たとえば，声道長の差異を近似する対数パワースペクトルの周波数ウォーピングは，ケプストラムに対する線形変換で表されることが示されている[10)]．また，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に対する区分的な線形変換と仮定し，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[11)]．さらに，音声認識における音響モデルの MLLR（Maximum Likelihood Linear Regression）適応でも，話者の違いをケプストラムの区分的な線形変換と仮定し，パラメータを最尤推定することで実現されている．以上のことから，理論的にも経験的にも，話者の違いはケプストラムに対する可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ で近似できることが分かる．

次に，声質と同様に不可避な非言語的特徴として，録音機器や伝送経路の周波数特性の違いがある．音声をコンピュータで分析するためには，必ず音声を電気的信号に変換してコンピュータに送らなければならないため，この非言語的特徴は不可避である．周波数特性の違いは，パワースペクトルに対する乗算で表現されるため，ケプストラム領域では定ベクトルの足し算で表現される．足し算は当然可逆な変換であるので，話者性の違いと録音機器の違いは，両方含めてケプストラムに対する可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ で近似されることになる．

不可避的ではないが多くの場合に含まれてしまう非言語的特徴として，背景雑音の違いがある．背景雑音は，静かな部屋に移動したり，指向性の高いマイクを用いて音声を収録したりすることなどで，ある程度低減することができるが，現実的にはある程度の背景雑音は含まれてしまう．背景雑音の違いは，パワースペクトルに対する足し算で表現されるため，ケプストラム領域では可逆な非線形変換になる．しかし，背景雑音は時間変動するため，変換関数も時間依存になる．ただし，時間に依存せず周波数特性が一定な背景雑音に関しては，声質や録音デバイスの違いと同じく，可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ で表現される．

2.2 音声の構造的表象

ここまで，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，可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ で近似できることを見てきた．そのため，このような変換の前後であまり変化しない特徴量は，多くの非言語的特徴に非常に頑健である

といえる．そこで，そのような特徴量について考えていく．

まずケプストラムが，ある単位（たとえば音素）ごとに，ある一定な分布から出力されているものと仮定する．これは，HMM を利用した音声処理でも仮定されていることである．このような仮定の下，ケプストラムの時系列 $\boldsymbol{c}_t$（$t = 1, \cdots, T$）が与えられたとき，それを出力している分布群を推定することが可能である．これは HMM でいえば，与えられた 1 つの発声に対して，出力確率分布を学習することに相当する．

次に，2 つの分布間の距離尺度の 1 つである f-divergence（f-div.）について考える．ケプストラムを出力している複数の分布が推定されれば，それらの分布の間の f-div. を計算することが可能である．2 つの分布 p_i，p_j 間の f-div. は以下の汎関数で定義される．

$$f_{div.}(p_i, p_j) = \int p_j(\boldsymbol{c})\, g\left(\frac{p_i(\boldsymbol{c})}{p_j(\boldsymbol{c})}\right) d\boldsymbol{c} \tag{1}$$

ただし，$g(x)$ は $x > 0$ で定義される凸関数であり，$g(x)$ を換えることで様々な f-div. が定義可能である．たとえば $g(t) = t\log(t)$ とすれば f-div. はカルバック・ライブラ距離，$g(t) = \sqrt{t}$ とすれば，$-\log(f\text{-div.})$ はバタチャリヤ距離になる．

ここで，可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ がケプストラム空間全体を連続的に写像しているとすると，ケプストラムを出力する分布間の f-div. は，このような変換 h に不変となる[12]．可逆なケプストラム空間の写像 $\boldsymbol{c}' = h(\boldsymbol{c})$ により，ケプストラムの出力分布 p_i，p_j がそれぞれ q_i，q_j に変換される場合の f-div. の不変性は，$J(\boldsymbol{c}')$ を $h^{-1}(\boldsymbol{c}')$ 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．

$$f_{div.}(p_i, p_j) = \int p_j(\boldsymbol{c}) g\left(\frac{p_i(\boldsymbol{c})}{p_j(\boldsymbol{c})}\right) d\boldsymbol{c} \tag{2}$$

$$= \int p_j(h^{-1}(\boldsymbol{c}')) g\left(\frac{p_i(h^{-1}(\boldsymbol{c}'))J(\boldsymbol{c}')}{p_j(h^{-1}(\boldsymbol{c}'))J(\boldsymbol{c}')}\right) J(\boldsymbol{c}') d\boldsymbol{c}' \tag{3}$$

$$= \int q_j(\boldsymbol{c}') g\left(\frac{q_i(\boldsymbol{c}')}{q_j(\boldsymbol{c}')}\right) d\boldsymbol{c}' = f_{div.}(q_i, q_j) \tag{4}$$

以上により，f-div. が可逆な変換に不変になる十分性が証明された．なお証明は省くが，連続かつ可逆な写像 $\boldsymbol{c}' = h(\boldsymbol{c})$ に不変な分布間距離尺度は f-div. 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[12]．

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，それらの分布すべての間の f-div. を計算することで，1 つの距離行列が得られる．距離行列は 1 つの幾何学的な形態を規定するため，これを，音声の構造的表象と呼んでいる．音声の構造的表象を構成する各エッジは f-div. であるので，音声の構造的表象は可逆変換 $\boldsymbol{c}'_t = h(\boldsymbol{c}_t)$ で表現できる非言語的特徴に理論的に不変となる．**図 1** に，ケプストラム空間において音素を音響イベントとする構造的表象が不変になる様子を示す．

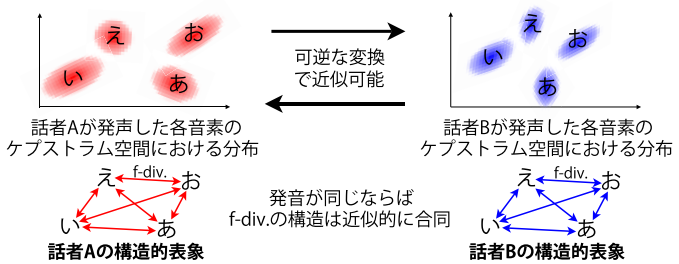


図 1　静的で可逆な変換に不変な音声の構造的表象
Fig. 1　Transform-invariant pronunciation structures.

なお本研究では f-div. として，すでに音声認識タスクで性能が高いことが示されているバタチャリヤ距離（Bhattacharyya Distance; BD）の平方根を構造的表象の抽出に利用する[7]．2 つの分布 p_i，p_j 間のバタチャリヤ距離 BD は，以下のように定義される．

$$BD(p_i, p_j) = -\log \int \sqrt{p_i(\boldsymbol{c})p_j(\boldsymbol{c})}d\boldsymbol{c} \tag{5}$$

なお，2 つの分布がガウス分布 $\mathcal{N}_i(\mu_i, \Sigma_i)$，$\mathcal{N}_j(\mu_j, \Sigma_j)$ だった場合，BD はガウス分布の平均と分散共分散行列の閉形式で記述できる．

$$BD\left(\mathcal{N}_i, \mathcal{N}_j\right) = \frac{1}{8}\left(\mu_i - \mu_j\right)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} \left(\mu_i - \mu_j\right) + \frac{1}{2}\log \frac{\left|\left(\Sigma_i + \Sigma_j\right)/2\right|}{\left|\Sigma_i\right|^{\frac{1}{2}} \left|\Sigma_j\right|^{\frac{1}{2}}} \tag{6}$$

2.3　音声の構造的表象を用いた外国語発音分析

従来の音声の構造的表象を用いて外国語発音評価を行う枠組みを，**図 2** に示す．まず，音声を音素などの単位で分布化し，それらの距離行列を計算することによって音声の構造的表象を抽出する．次に，学習者の発声から抽出した距離行列と，教師の発声から抽出した距離行列を比較し，どの要素にどの程度差異があるかを見て発音を評価する．

2 つの構造間差異 d には，学習者の距離行列を S，教師の距離行列を T として，

$$d = \sum_{i<j} D_{ij} = \sum_{i<j} \left(\frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}}\right)^2 \tag{7}$$

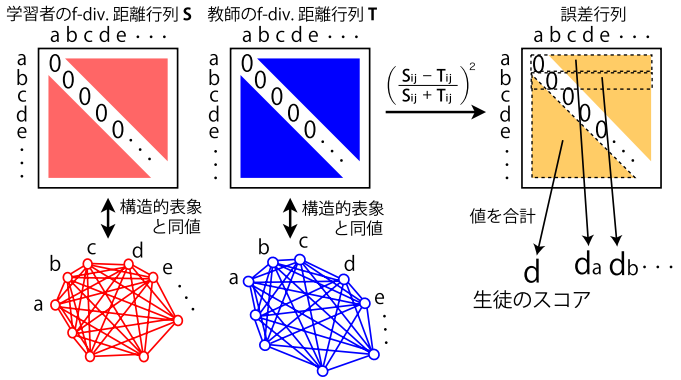


図 2　音声の構造的表象を用いた外国語発音評価
Fig. 2　Structure-based pronunciation assessment.

が利用できる[13]．ただし，S_{ij}，T_{ij} はそれぞれ S，T の i 行目 j 列目の要素，すなわち i 番目の分布と j 番目の分布の $\sqrt{BD}$（式 (5) の平方根）を表す．また，D は S と T の要素ごとの差の 2 乗を表す行列で，これを誤差行列と呼ぶ．

なお誤差行列は，上三角部分の和をとって学習者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（たとえば音素ごと）の評価も可能である．たとえば，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．

$$d_a = \sum_i D_{ai} = \sum_i \left(\frac{S_{ai} - T_{ai}}{S_{ai} + T_{ai}} \right)^2 \tag{8}$$

構造的表象は，あらゆる可逆な変換に対して不変である．そのため構造的表象は，CMN（Cepstrum Mean Normalization），最適な SS（Spectral Subtraction），最適な VTLN（Vocal Tract Length Normalization），最適な大局的 MLLR 適応などにも不変であり，それらを明示的に行わなくても，等価な効果を得られることが期待できる．

3. 構造的表象と多段階の重回帰分析を用いた発音分析

3.1 2 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，音響イベント数 M に対して ${}_MC_2$ 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の次元は ${}_{42}C_2 = 861$ 次元となってしまう．次元数が高くなると，発音分析に無関係な次元が増え，球面集中現象と呼ばれる現象により特徴量空間の 2 点間距離がほぼ一定になり，パターン認識問題の精度が極端に低下してしまう「次元の呪い」の問題が発生してしまい，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．実際に人間が発音評価を行うときにも，似たような処理を行っていると考えられる．たとえば，音声学者がある母音の発声についてその適切さを評価する場合には，その母音と一部の他母音との区別が十分実現されているかについて注意を払うことが知られている[14]．これは，発音評価の際に参照するエッジの種類が部分的であることを示唆しており，次元圧縮と同等の処理を行っていると考えられる．

ところが今回のタスクでは，単純な線形変換による次元圧縮を用いるには問題がある．861 次元の特徴量を次元圧縮するためには，最低でも 861 個の学習データが必要になり，十分な精度で次元圧縮を行うには非常に多くの学習データが必要となってしまう．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，次元圧縮後の特徴量の解釈が難しくなってしまうという問題もある．誤差行列は，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，次元圧縮を用いることで，その情報が消えてしまう．

そこで，少ないデータで学習可能であり，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，2 段階重回帰分析を提案する．音声の構造的表象と 2 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．2 段階重回帰では，誤差行列を 2 段階に分けて重回帰分析を行う．2 段階で重回帰分析を行うことにより，学習する重み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，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．各音素ごとの手動評価値が得られない場合には，学習者に対する手動評価値を利用することもできる．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．たとえば，日本人が発音する米語の/s/を評価する

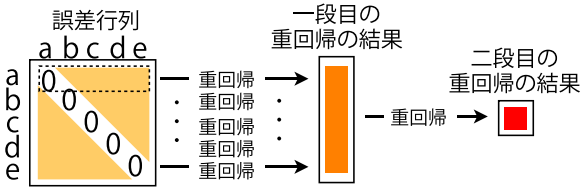


図 3　2 段階重回帰分析を用いた外国語発音評価
Fig. 3　Two-layered regression analysis.

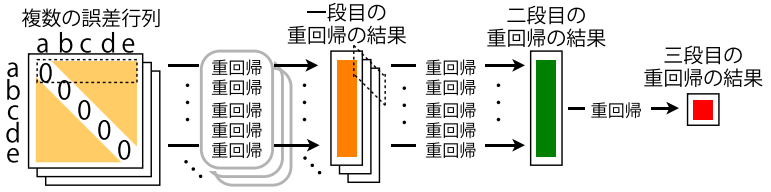


図 4　3 段階重回帰分析を用いた外国語発音評価
Fig. 4　Three-layered regression analysis.

際には，日本人が/s/と混同しやすい/ʃ/や/θ/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，各音響イベントごとの評価値が得られる．

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，学習者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，学習者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．たとえば，日本人の米語発音を評価する場合には，日本語の/え/の発音で置き換えてほぼ問題ない/ɛ/のような音素は無視され，/ɚ/など，発音の上手/下手の差が大きい音素が重要視されると予想される．2 段目の重回帰の結果，学習者の評価値が得られる．

3.2　複数の誤差行列と 3 段階重回帰分析

ここまでは，学習者と教師の構造的表象を 1 対 1 で比較した誤差行列のみを使って発音分析を行っていた．そのため，教師の方言の違いや発音の癖なども含めて，学習者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，学習者と複数の教師の構造的表象を 1 対多で比較すると，複数の誤差行列が得られる．複数の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だと考えられる．

また教師を複数用いるほかにも，音響特徴量を変更すれば，誤差行列を複数得ることができる．たとえば，音声分析には多くの場合 16 kHz 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 3 kHz 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[15]．ここで，通常の 16 kHz サンプリングの音声と，6 kHz サンプリングの音声を用意し，それぞれから MFCC を計算すれば，2 つの音声の構造的表象および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，誤差行列を複数得ることが可能である．これにより，情報量が増えるという利点がある．ただしそれと同時に，次元数が誤差行列の数だけ倍に増えるため，次元の呪いの問題がさらに強くなってしまうという欠点がある．

そこで，2 段階重回帰分析を拡張した，3 段階重回帰分析により，複数の誤差行列から次元の呪いを適切に避け，情報量が増えるという利点を生かす手法を提案する．3 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 2 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各ストリーム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，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し，それが得られない場合は話者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，どの音響特徴量空間の，どの教師との誤差行列を重要視すべきかが学習される．たとえば，16 kHz サンプリング音声の MFCC と 6 kHz サンプリング音声の MFCC から計算した 2 つの誤差行列を用いる場合，3 kHz 以下に音素の特徴が多く含まれる母音は 6 kHz サンプリングの構造が重要視され，高周波数領域にも音素の特徴が多く含まれる子音は 16 kHz 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合せ

音声の構造的表象を用いた発音評価では，音の絶対的な特徴を捨て，音と音との相対関係のみで発音を記述している．一方 HMM を用いた発音評価手法では，HMM でモデル化された音そのものの特徴で発音を記述している．そのため，両者は音声の異なる特徴をとらえており，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．たとえば，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

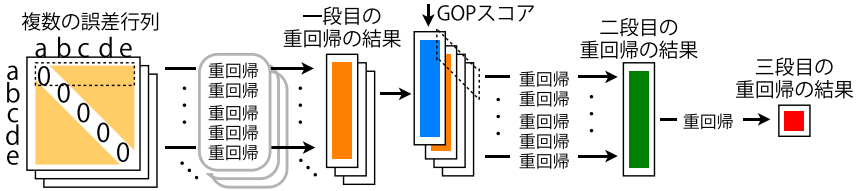


図 5　3 段階重回帰分析と GOP スコアを組み合わせた外国語発音評価
Fig. 5　Three-layered regression analysis with GOP scores.

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 1 つとして，GOP（Goodness Of Pronunciation）スコアを採用する[16]．GOP スコアは，以下の近似式を用いて計算される．

$$GOP(c_1,\ldots,c_T,p_1,\ldots,p_N) = \log P(p_1,\ldots,p_N|c_1,\ldots,c_T) \tag{9}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\frac{1}{D_{p_n}}\log\left\{\frac{P(c^{p_n}|p_n)}{\sum_{q\in Q}P(c^{p_n}|q)}\right\} \tag{10}$$

$$\approx \frac{1}{N}\sum_{n=1}^{N}\frac{1}{D_{p_n}}\log\left\{\frac{P(c^{p_n}|p_n)}{\max_{q\in Q}P(c^{p_n}|q)}\right\} \tag{11}$$

ここで，p_n は音素，T は観測フレーム長，N は意図された音素数，c^{p_n} は強制アライメントによって得られた音素 p_n に対応する音声区間，D_{p_n} はその区間長，Q は全音素集合である．なお $\{c^{p_1},\cdots,c^{p_N}\}$ と $\{c_1,\cdots,c_T\}$ は同一の音響特徴量時系列を表現している．式 (11) のサメンションの中の式は，各音素ごとの GOP スコアとなる．

各音素ごとの GOP スコアを 3 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や，HMM の適応時に過学習が起こりにくい音素では 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．

表 1　音声の構造的表象を抽出するための音響分析条件
Table 1　Acoustic analysis condition for pronunciation structures.

サンプリング	16 bit / 16 kHz
窓	25 msec 幅，10 msec シフトのハミング窓
学習データ	1 名につき 60 文（set 8 を読み上げた話者のみ 40 文）
HMM 学習時の特徴量	MFCC(12) + Energy(1) + ΔMFCC(12) + ΔEnergy(1)
構造抽出時の特徴量	MFCC(12)
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	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

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験を行った[17]．ERJ データベースは，200 名の日本人米語学習者（大学生）と米語母語話者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 TIMIT に含まれる 60 文などが含まれている．200 名の学習者それぞれは，この文セットのうち 1 セットのみを読み上げている．そのため 1 セットにつき約 25 名の学習者が読み上げている計算になる．また，20 名の米語母語話者それぞれが，文セットのうち 4 セットを読み上げている．ただし，20 名中 2 名（男性 M08 と女性 F12）は，全 8 セットを読み上げている．さらに，各学習者が読み上げた文章のうち，1 人につき 10 文の発声に対して手動評価値が付与されている．評価者は日本人学習者の癖をよく理解している，米語母語話者である音声学者 5 名である．評価は，音素的に正しく発声されたか，リズムが正しく発声されたか，イントネーションが正しく発声されたか，の 3 つの尺度が用いられ，それぞれ 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，学習者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 $\sqrt{BD}$ 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 1 つの分布になっているが，2 つの音素 HMM の対応する 3 つの分布間の $\sqrt{BD}$ の和を 1 つのエッジとして構造的表象を作成する．これにより，音響イベントが音素になり，${}_{42}C_2 = 861$ 本のエッジから

表 2　GOP 算出に用いる HMM を学習するための音響分析条件
Table 2　Acoustic analysis condition for GOP scores.

サンプリング	16 bit / 16 kHz
窓	25 msec 幅，10 msec シフトのハミング窓
学習データ	ERJ に含まれる 20 名の母語話者の音声すべて
話者適応用データ	60 文（set 8 を読み上げた話者のみ 40 文）
特徴量	CMN をかけた MFCC(12) + ΔMFCC(12) + ΔEnergy(1)
HMM	不特定話者音素 monophone HMM
出力確率分布	対角共分散行列を持つ 4 混合 GMM
トポロジ	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

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．

このように抽出した構造的表象を用いて，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．ここで手動評価値には，ERJ の各話者につけられた 10 文 × 5 名の音素的正しさに対する評価値の平均値を利用した．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としてこの話者の手動評価値を用いた．また 2 段階重回帰分析における教師の音声には，男声教師（M08）1 名のみを利用した．3 段階重回帰分析の際には，教師の音声に M08 と F12 の 2 名を，さらに表 1 の条件の MFCC と，3 kHz ローパスフィルタを通した音声の MFCC の 2 種類の音響特徴量を用い，$2 \times 2 = 4$ の誤差行列を用いた．また GOP スコアを算出するための HMM の学習には，**表 2** の条件を用い，大局的な MLLR 適応をかけて利用した．なお，回帰木などを用いた通常の MLLR 適応を用いなかった理由は，回帰木のノード数を増やすことで発音評価の精度が逆に低下することが知られているためである[4]．

なお，今回の実験条件における各段階の重回帰分析は，約 175 の学習データを用いて約 40 の重みパラメータを推定する問題になるので，過学習が発生する恐れがある．そこで，通常の最小誤差基準に，2 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ることにより，汎化性能を向上させパラメータ推定の信頼度を向上させた．リッジ回帰に用いる正則化パラメータ λ は，すべて 1 とした（ただし 3 段階重回帰分析の 2 段目の重回帰分析は，学習データが約 175，推定するパラメータが 4 であるため，$\lambda = 0$ として，通常の最小二乗誤差基準の重回帰分析を行った）．

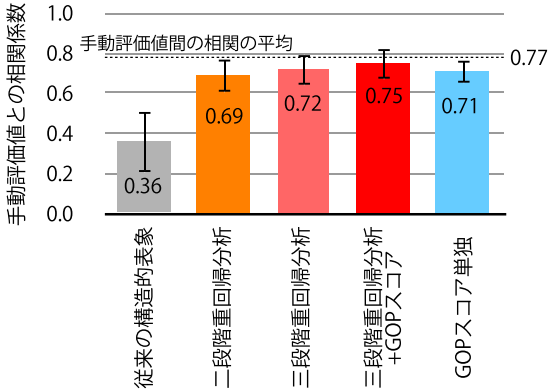


図 6　手動評価値との相関の平均値と標準偏差
Fig. 6　Averages and standard deviations of correlation coefficients between human and machine.

さらに比較実験として，多段階重回帰分析を用いない従来の構造的表象を用いた実験と，GOP スコアを単独で用い，音素ごとの GOP スコアを重回帰分析することで評価値を算出する実験も行った．

結果を**図 6** に示す．結果，多段階重回帰分析を用いることで，従来の構造的表象を用いた手法から大幅に相関が向上していることが分かる．また，2 段階重回帰分析より，3 段階重回帰を用いた方がわずかに相関が向上し，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分かる．また，GOP スコアを重回帰分析した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．なお，各手動評価者による評価値間の相関の平均は 0.77 となっており，提案手法は手動評価者とほぼ同等の精度で発音評価が行えているということができる．

なお，詳細な分析は行っていないが，多段階重回帰分析で学習された重みパラメータは，たとえば/l/と/r/間のエッジの重みが非常に大きくなっているなど，その解釈が比較的容易な数値となっている．

4.2　ミスマッチ条件下での実験

構造的表象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学習者の読み上げ音声に人工的な声道長変換をかけた音声の発音分析実験を行った．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

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{12}$$

による近似[18]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[19]．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．ただし，$|\alpha| < 1$であり，$\alpha = 0$のときに変換なし，$\alpha = \pm 0.40$のときに，声道長が約半分/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2段階重回帰分析，3段階重回帰分析，3段階重回帰分析+GOPスコアを用いた実験と，比較実験として従来の構造的表象を用いた手法，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*1．まず構造的表象を用いた手法はすべて，声道長を変化させてもほとんど相関は変化しておらず，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．先の実験と同様，GOPスコアを組み合わせた3段階重回帰分析が，どのウォーピングパラメータにおいても最も高い性能を示している．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

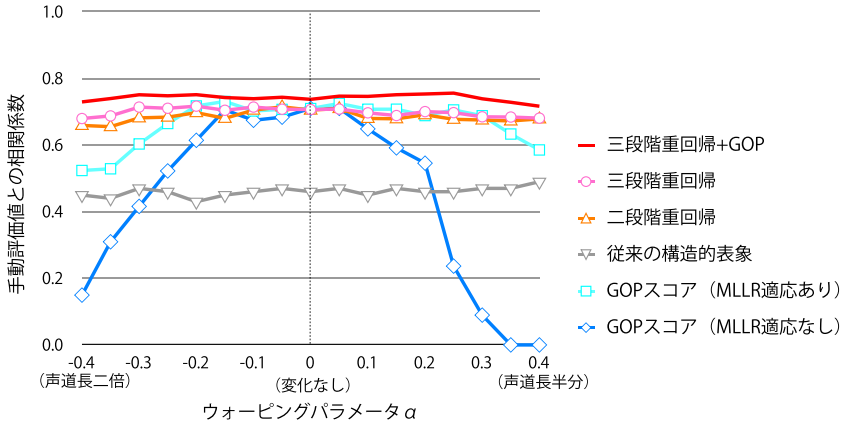


図7　声道長変換をかけた音声の手動評価値との相関の平均値
Fig. 7　Averages of correlation coefficients between human and machine with warped utterances.

*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．

幅に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，提案手法は大局的な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分かる．

4.3　音素ごとのスコアの推定実験

次に，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．音声データとして，日本人中学生36名（平均13.5歳）と5名の英語教師（平均45.2歳）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．録音は，通常の教室で，ヘッドセットを用い16 bit/16 kHzサンプリングで録音した．この音声データを，音声学者3名が，各母音それぞれのスコアおよび各学習者ごとの総合スコアを4段階評価した．また，3名のうち1名は，時期をずらして評価を2回行った．なお，この評価値の統計量として，異なる3名の音声学者の手動評価値の相関（評価者間相関）の平均値を**表3**に，1名の音声学者が2回評価したときの評価値の相関（評価者内相関）を**表4**に示す．また，手動評価値の平均値および標準偏差を，**表5**に示す．表3，表4でともに相関が低い母音（/ɛ/，/uː/など）は，日本語母語話者にとっては発音するのが容易な母音であるため，各話者で評価値の差が小さく，評価値の平均が高く，標準偏差が低くなっていることが分かる．

実験はまず，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，構造的表象を抽出した．なお分布化の際には，少ない

表3　各手動評価者の手動評価者間の相関の平均
Table 3　Averages over inter-rater correlations.

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	0.45

表4　1名の手動評価値間の相関
Table 4　Intra-rater correlations.

ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː	総合
0.79	0.85	0.37	0.61	0.33	0.82	0.73	0.84	0.59	0.49	0.63

表5　手動評価値の平均と標準偏差
Table 5　Averages and standard deviations of scores.

	ɑː	æ	ʌ	ɔː	ɛ	ɚː	ɪ	iː	ʊ	uː	総合
平均	2.40	2.37	3.00	2.60	3.39	2.24	2.55	3.35	3.04	2.56	2.48
標準偏差	0.87	0.95	0.70	0.64	0.36	1.13	0.93	0.84	0.58	0.57	0.55

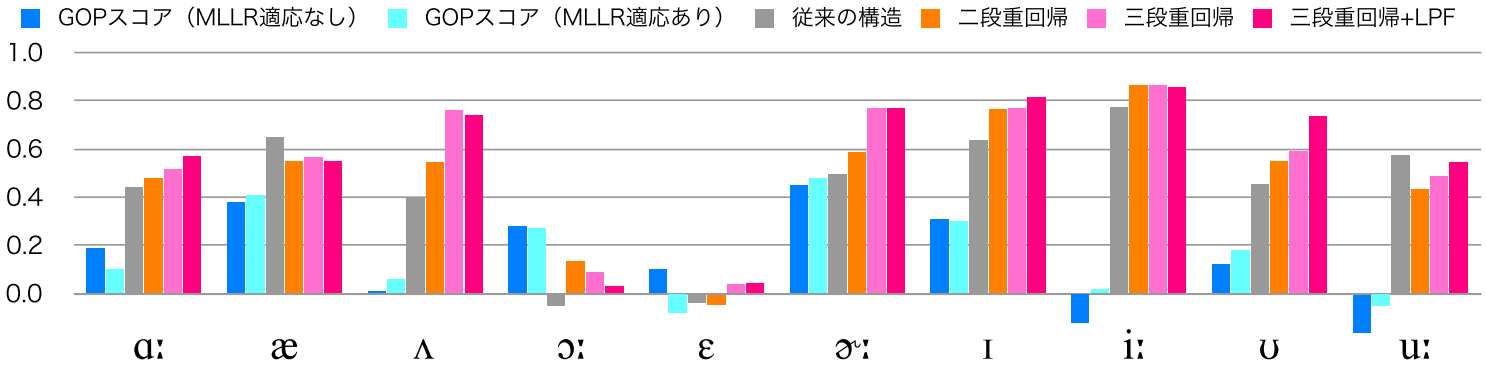


図 8　母音ごとの手動評価値との相関係数
Fig. 8　Correlations of the vowel-based scores between human and machine.

表 6　総合評価値の手動評価値との相関係数
Table 6　Correlations of the overall pronunciation proficiency.

GOP（適応なし）	GOP（適応あり）	従来の構造	2 段階	3 段階	3 段階+LPF
0.45	0.28	0.76	0.79	0.87	0.88

データから信頼できる分布を得るために，MAP 推定を用いた[6]．重回帰分析には，先の実験と同様リッジ回帰分析を用いた．2 段階重回帰分析では，教師の構造には男声話者 1 名分のみを利用した．3 段階重回帰分析では，7 名の教師から誤差行列を利用したものと，同様に 7 名の教師と通常の MFCC と 3 kHz ローパスフィルタ（LPF）をかけた MFCC を用いて $7 \times 2 = 14$ の誤差行列を利用するもの，2 通りを実験した．さらに比較実験として，従来の構造的表象を用いた実験と，GOP スコアを用いて音素ごとのスコアを算出し，それらを重回帰分析することで話者ごとのスコアを算出する実験も行った．なお，GOP スコア算出に用いる HMM には，前節と同じ物を利用した．

まず，話者ごとの自動評価値と手動評価値の総合スコアとの相関を**表 6** に示す．GOP を用いる手法と構造的表象を用いる手法を比較すると，構造的表象を用いる方が大幅に相関が高いことが分かる．今回用いた音声データは，中学生の発音データで，かつ 10 母音の 1 回ずつの発声のみと，1 人あたりの発声数が非常に少なく偏っている．そのため MLLR 適応なしの場合にはミスマッチ問題が発生し，MLLR 適応ありの場合には過適応の問題が発生しているものと考えられる．次に，構造的表象を用いる手法の中で比較すると，提案手法である多段階重回帰分析を用いることで，相関はさらに向上している．ただし従来の構造的表象を用いる手法でも，十分高い相関が得られている．これは，10 母音から構造的表象を抽出しているため，構造のエッジの数は 45 本と比較的少なく，次元の呪いの問題が発生しないためと考えられる．また，3 段階重回帰分析において，LPF をかけたものを加えると加えないでは，ほとんど性能の変化が見られない．このことは，LPF をかけていなくても構造的表象が十分に非言語情報をキャンセルできていることを示唆している．

次に，母音ごとの自動評価値と，各母音の手動評価値との相関を**図 8** に示す．まず，GOP スコアを用いる手法と構造的表象を用いる手法を比較すると，/ɔː/を除き，構造的表象を用いる手法の方が相関が高いことが分かる．構造的表象を用いる手法の中では，表 5 に示した手動評価値の標準偏差が大きい母音（/ɑː/，/æ/，/ɚː/，/ɪ/，/iː/），すなわち話者によって上手/下手のレベルの差がつきやすい母音については，/æ/を除き，多段階重回帰分析を行うことで性能が向上していることが分かる．また，他の 5 母音においても，/ʌ/，/ʊ/については提案手法の有効性が確認できる．/ɛ/については，どの手法でも相関値は非常に低くなっているが，表 5 における標準偏差も非常に低いので，この結果は比較的妥当だと考えられる．/ɔː/については，構造的表象を用いる方法が評価に不十分であるといえる．これは，bought（[ɔː]）の発音を [oʊ] と発音間違いする話者が多く，構造的表象が二重母音の動きを適切にとらえられていないために性能が低下してしまったものと考えられる．

5. おわりに

本稿では，構造的表象を用いた外国語発音評価において，多段階重回帰を提案した．多段

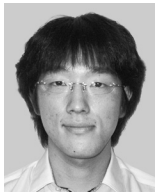
階重回帰を用いることで，適切な自由度と汎化性能を持つ次元圧縮により精度の高い発音評価が可能になり，さらに，途中段階で音素ごとの評価値を得ることができる．また，GOP スコアと構造的表象を適切に組み合わせることも可能になる．母音子音すべて含む音声を用いて発音評価実験を行った結果，従来の構造的表象を用いた外国語発音評価手法に対し，提案手法により大幅に精度が向上することが分かった．また，GOP スコアを利用した手法と比較しても，提案手法はより高い精度が得られた．さらに GOP と提案手法を組み合わせることで，より高い精度が得られた．さらに，少ないデータから母音のみの発音評価実験を行った結果，提案手法は GOP スコアを用いる手法より大幅に高い精度が得られることが分かった．

今後の課題としてはまず，様々な声道形状を持つ話者の実音声を用いた，母音子音すべてを含む場合での発音評価実験を行うことがあげられる．次に，より少ない音声から構造的表象を抽出する手法の開発があげられる．今回の実験では，母音子音すべてを評価する場合には 50 文章分の読み上げ音声を利用しているが，これを 1 文章からでも評価できるようにし，教育現場で用いやすくしたい．他には，発音分析の高精度化があげられる．本稿では，構造的表象を用いた発音評価のみを扱ったが，構造的表象を用いることで，学習者の発音分類など，発音分析を行うことも可能である[6),13)]．これらの発音分析に関しても，適切なエッジの重み付けにより高精度化させていきたい．

参 考 文 献

1) 鳥飼玖美子：危うし！小学校英語，文藝春秋 (2006).
2) Russell, M. and D'Arcy, S.: Challenges for computer recognition of children's speech, *Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education* (2007).
3) Ohkawa, Y., Suzuki, M., Ogasawara, H., Ito, A. and Makino, S.: 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).
4) Luo, D., Qiao, Y., Minematsu, N. and Hirose, K.: Regularized maximum likelihood linear regression adaptation for computer-assisted language learning systems, *IEICE Trans. Information and Systems*, Vol.E94-D, No.2, pp.308–316 (2011).
5) Minematsu, N.: Yet another acoustic representation of speech sounds, *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, pp.585–588 (2004).
6) 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).
7) Qiao, Y., Asakawa, S., Minematsu, N. and Hirose, K.: On invariant structural representation for speech recognition: Theoretical validation and experimental improvement, *Proc. INTERSPEECH*, pp.3055–3058 (2009).
8) Minematsu, N., Qiao, Y., Asakawa, S. and Suzuki, M.: Speech structure and its application to robust speech processing, *Journal of New Generation Computing*, Vol.28, No.3, pp.299–319 (2010).
9) 朝川　智，喬　宇，峯松信明，広瀬啓吉：音声の構造的表象と判別分析を用いた単語音声認識，電子情報通信学会技術研究報告，SP2008-113, pp.203–208 (2008).
10) Pitz, M. and Ney, H.: Vocal tract normalization equals linear transformation in cepstral space, *IEEE Trans. Speech and Audio Processing*, Vol.13, No.5, pp.930–944 (2005).
11) 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).
12) Qiao, Y. and Minematsu, N.: A study on invariance of f-divergence and its application to speech recognition, *IEEE Trans. Signal Processing*, Vol.58, No.7, pp.3884–3890 (2010).
13) Suzuki, M., Minematsu, N., Luo, D. and Hirose, K.: Sub-structure-based estimation of pronunciation proficiency and classification of learners, *Proc. Int. Workshop on Automatic Speech Recognition and Understanding*, pp.574–579 (2009).
14) 鎌田　圭，朝川　智，峯松信明，牧野武彦，広瀬啓吉：音声の構造的表象を用いた英語学習話者の分類に関する実験的検討，電子情報通信学会技術研究報告，SP2006-77, pp.7–12 (2006).
15) Kitamura, T., Honda, K. and Takemoto, H.: Individual variation of the hypopharyngeal cavities and its acoustic effects, *Acoustical Science and Technology*, Vol.26, No.1, pp.16–26 (2005).
16) Witt, S. and Young, S.: Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning, *Speech Communication*, Vol.30, pp.95–108 (2000).
17) 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).
18) 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).
19) Kawahara, H.: STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds, *Acoustical Science and Technology*, Vol.27, No.6, pp.349–353 (2006).

(平成 22 年 10 月 14 日受付)
(平成 23 年 2 月 4 日採録)



鈴木　雅之（学生会員）

2010 年東京大学大学院工学系研究科修士課程修了．修士（工学）．現在，同大学院工学系研究科博士後期課程に在籍．音声認識，音声分析，音声強調に関する研究に従事．IEEE，ISCA，電子情報通信学会，日本音響学会各会員．

峯松　信明（正会員）

1995 年東京大学大学院工学系研究科博士課程修了．博士（工学）．同年豊橋技術科学大学情報工学系助手．2000 年東京大学大学院工学系研究科助教授．2002 年瑞国王立工科大学客員研究員．現在，東京大学大学院情報理工学系研究科准教授．音声科学から音声工学に至るまで幅広く音声コミュニケーションに関する研究に従事．IEEE，ISCA，IPA，CALICO，電子情報通信学会，日本音響学会，人工知能学会，日本音声学会，日本音声言語医学会，外国語教育メディア学会等各会員．

広瀬　啓吉（正会員）

1972 年東京大学工学部電気工学科卒業．1977 年同大学大学院博士課程修了．工学博士．同年東京大学工学部電気工学科講師．1994 年同電子工学科教授．1996 年東京大学大学院工学系研究科電子情報工学専攻教授．1999 年同新領域創成科学研究科教授．2004 年 10 月より同情報理工学系研究科教授．1987 年米国 MIT 客員研究員．音声言語情報処理分野一般についての研究開発に従事，特に韻律に着目した研究．IEEE，米国音響学会，ISCA（Board メンバ），電子情報通信学会（フェロー），日本音響学会，人工知能学会，言語処理学会，信号処理学会各会員．