

## 構造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, 喬宇, 峯松信明, 広瀬啓吉 (東大), 牧野武彦 (中央大)

### 1 はじめに

音声は環境により大きく変形する。そのため、音響モデルを構築する際どんなに多くのデータを集めても、モデルと入力とのミスマッチ問題を避けることができない。ミスマッチ問題の解決法として、音響モデルの適応が広く行われているが、“There is no data like more data.”の言葉通りデータは本質的に不足しており、汎化能力の高いなんらかの手法による解決が求められる。例えば Furui は音声認識のレビュー論文において、HMM を識別学習・適応する際、汎化能力を高めるさまざまな工夫をした手法を紹介すると共に、ゴールデン・スタンダードと呼べる汎化手法はまだ存在しないと述べている [1]。

外国語発音評価システム用音響モデルの学習・適応では、汎化能力を高めることはより難しい問題になる。例えば、母語や発音習熟度によっても音声を変形するため、過適応の問題が発生しやすくなる [2]。また、2011 年度からの小学校 5・6 年生の英語活動必修化を受けて学習者に小さな子供が増えることが予想されるが、子供の音声は大人の音声よりも変形が大きく、汎化はより難しい [3]。

我々は外国語発音評価において、解くべき課題の多い HMM の適応とはまったく異なるミスマッチ問題の解決法として、静的な変形に不変な音声の相対関係の利用を提案している [4]。これは、音声の変形に近似的に不変な特徴量を用いることでミスマッチを低減させる手法であり、HMM の適応とは異なる視点からのアプローチとなる。我々はこの音声の相対関係から得られる情報表象を音声の「構造表象」と呼び、外国語発音評価における有効性を示してきた [5, 6]。また、構造表象を利用した音声認識も検討し、構造表象による音声分析技術を高度化させてきた [7, 8]。

本研究では、構造表象を用いた外国語発音評価において、適当な汎化能力を持つ次元圧縮を用いて精度を改善する手法を提案する。提案手法により、従来の構造表象を用いた外国語発音評価手法から大幅に精度が向上した。また、外国語発音評価に広く用いられている、HMM の事後確率を用いた GOP (Goodness Of Pronunciation) スコア [9] 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 スコアと提案手法を組み合わせることで、さらに高い精度が得られることも示す。

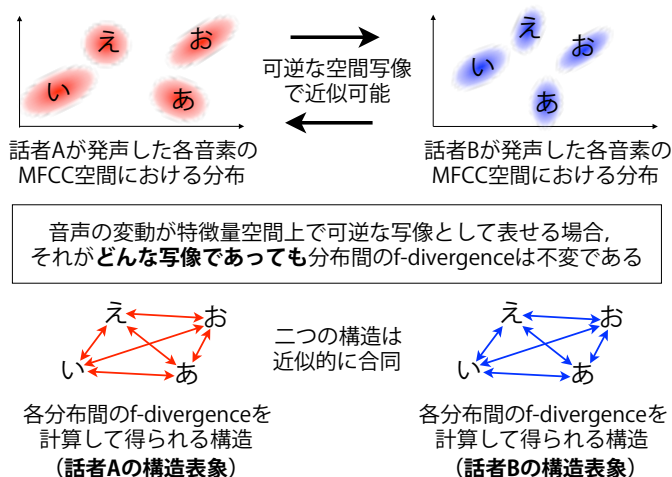


Fig. 1 音声の静的な変形に不変な構造表象

### 2 音声の構造表象

話者の違いやマイクなどの伝送特性の違いは、ケプストラム空間における可逆な空間写像でモデル化できる。例えば MLLR 適応では、この空間写像を線形変換と仮定し、適切な変換パラメータを推定することによって実現される。また、GMM を用いた話者変換では、話者変換に対応する非線形写像を GMM で学習することによって実現している。

二つの空間が可逆な空間写像で結びつけられ、それぞれの空間において対応する複数の分布が存在する場合、それぞれの空間における分布間の  $f$ -divergence は、常に不変となる [10]。  $f$ -divergence とは、分布間距離尺度の一種であり、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。

$$f_{\text{div}}(p_i, p_j) = \int p_j(\mathbf{x}) g\left(\frac{p_i(\mathbf{x})}{p_j(\mathbf{x})}\right) d\mathbf{x} \quad (1)$$

Fig.1 に、MFCC 空間における対応する音響イベントの分布間の  $f$ -divergence が不変になる様子を示す。ケプストラム空間上で分布として表現された音響イベント間の  $f$ -divergence は、話者の違いや伝送特性などに近似的に不変になる。我々は、すべての音響イベント間の  $f$ -divergence を計算することによって得られる構造を、音声の構造表象と呼んでいる。

本研究では、構造表象を計算する実装において、  $f$ -divergence (の関数) として、Bhattacharyya Distance (BD) の平方根を構造の各エッジとして使用し

\* Automatic estimation of pronunciation proficiency based on speech structure and multilayer regression.  
by M.Suzuki, Y.Qiao, N.Minematsu, K.Hirose (The Univ. of Tokyo), T.Makino (Chuo University)

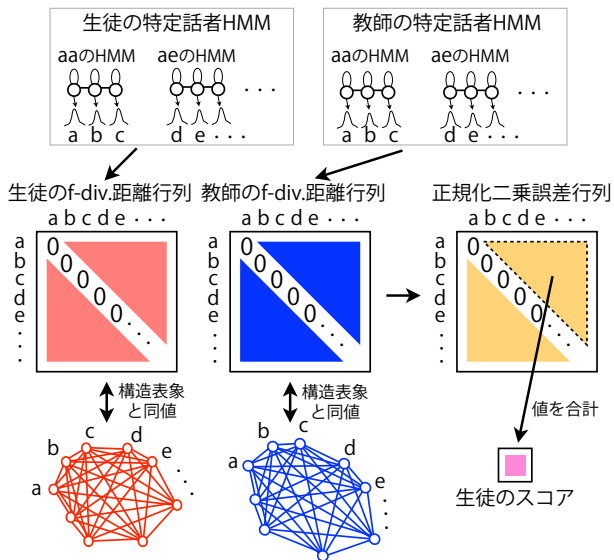


Fig. 2 構造表象を用いた外国語発音評価

ている. 二つの分布  $p_i(\mathbf{x}), p_j(\mathbf{x})$  間の BD は下記で定義される.

$$BD(p_i, p_j) = -\ln \int \sqrt{p_i(\mathbf{x})p_j(\mathbf{x})} d\mathbf{x} \quad (2)$$

### 2.1 構造表象を用いた外国語発音評価

構造表象を用いた外国語発音評価システムを実装するには, 生徒の音声, 教師の音声からそれぞれ構造表象を抽出し, 生徒の構造表象と教師の構造表象の構造表象間差異が小さければよい評価を, 差異が大きければ悪い評価を与えればよい. システムを実装する上で問題になるのは, 1) どう音声を分布化するか, 2) どう構造表象間差異を定義するかの二つである.

音声を分布化する方法として, 今回は読み上げ文章から特定話者音素 HMM を学習する方法を用いる. 特定話者音素 HMM を学習すると, 各音素 HMM の状態ごとに出力確率分布が得られる. これを一つの音響イベントとして利用する. このようにして音声音が音響イベント単位で分布化されれば, すべての分布ペアの  $\sqrt{BD}$  距離行列を計算することで構造表象が得られる.

生徒と教師の構造表象間の差異尺度としては, [5] で提案されている構造の各エッジごとの正規化二乗誤差の合計が利用できる. 生徒の構造  $S$  と教師の構造  $T$  の正規化二乗誤差の合計  $D$  は, 以下のようにかける.

$$D(S, T) = \sum_{i < j} \left( \frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}} \right)^2 \quad (3)$$

ここで,  $S_{ij}, T_{ij}$  は, それぞれ生徒と教師の音響イベント分布  $p_i, p_j$  間の  $\sqrt{BD}$  である.  $\sqrt{BD}$  は対称性のある距離尺度であるので, 距離行列は対称行列となるため, 上三角成分のみを合計すればよい.

Fig. 2 に, 生徒と教師の特定話者音素 HMM から構造表象をそれぞれ作成し, 正規化二乗誤差の合計を計算することにより生徒のスコアを計算する一連の流れを示す. なお, 正規化二乗誤差行列は, 生徒と教師の発音がどう違うのかといった情報も含んでおり, 例えば各行の和を使えば, 生徒の発音をどの音から直していくべきかの推定が可能になる [6].

## 3 多段階重回帰を用いた外国語発音評価

本節から, 本研究における提案手法について述べる. 提案手法は, Fig. 2 において, 正規化二乗誤差行列を計算した後の処理に関する改善手法である.

### 3.1 構造表象と 2 段階重回帰

構造表象を表現するパラメータ数, すなわち構造のエッジの数は, 音響イベント分布の数を  $N$  個として,  $N(N-1)/2$  個となる. そのため, 音響イベントの数の二乗オーダーで次元数が増加してしまい, 次元の呪いが発生してしまう. この問題は特徴量の冗長性に起因しているため, 適当な次元圧縮法を用いることで解決することができる.

しかし, 教師あり次元圧縮法を単純に用いると, 次元数があまりに高すぎるために過学習がおき, 汎化性能が低下する問題が発生する. そのため, 構造表象を用いた音声認識タスクにおいては, エッジをランダムに選択してから判別分析する方法 [7] や, 2 段階で LDA をかける方法 [8] など, 学習するパラメータの数を少なくすることにより汎化能力を高めた次元圧縮手法が提案されている.

構造表象を用いた外国語発音評価タスクにおいては, 次元圧縮法として特徴量選択による部分構造化が提案されている [5]. PCA や LDA のような線形変換による次元圧縮を行っていない理由は, 外国語発音評定タスクでは正規化二乗誤差行列の各行の和なども情報として有用であるため, 正規化二乗誤差行列の形を崩さない次元圧縮法が有用であるためである. なお特徴量選択は, それぞれのエッジに対し, 使う/使わないの 2 値のラベルをふるものであり, 次元圧縮法としては最も自由度の低い (すなわち, 汎化能力は高いが性能は低い) 手法と言える.

これに対して本論文では, 部分構造化よりも自由度が汎化能力の低下を招かない程度に高く, かつ正規化二乗誤差行列の形も大きく崩さず意味解釈の行いやすい次元圧縮手法として, 2 段階重回帰を提案する. 2 段階重回帰を用いた外国語発音評価の枠組みを, Fig. 3 に示す.

1 段目の重回帰では, 正規化二乗誤差行列の行ごとに重回帰を行う. 重回帰分析の目的変数としては, 例

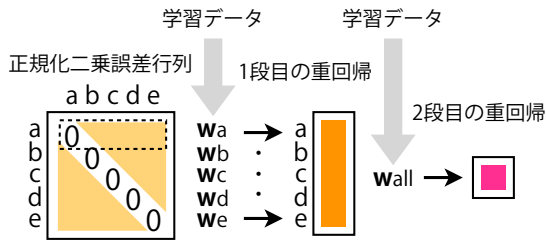


Fig. 3 2段階の重回帰を用いた外国語発音評価

えば各音素に対する手動評価値が利用できる。重回帰の重みパラメータ学習時には、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/er/を評価する際には、/er/と混同しやすい/aa/や/ae/などとの相対関係が重要視されると予想される。1段階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段階目の重回帰では、1段階目の重回帰で計算した各音響イベントの評価値に対して重回帰を行う。重回帰分析の目的変数としては、例えば生徒に対する手動評価値が利用できる。重回帰の重みパラメータの学習時には、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/s/などが重要視されると予想される。2段階目の重回帰の結果、生徒のスコアが得られる。

### 3.2 マルチストリーム構造と3段階重回帰

次に、マルチストリーム構造化と、2段階重回帰を拡張した3段階重回帰分析を用いたさらなる精度向上法について述べる。

まず、マルチストリーム構造化について述べる。構造表象は、例えばMFCC空間における分布間 $\sqrt{BD}$ を計算することにより算出されるが、例えば $\Delta MFCC$ 空間における分布間 $\sqrt{BD}$ を計算することによっても算出することができる。つまり、異なる音響特徴量を用いれば、異なる構造表象を抽出することができる。複数の音響特徴量を用いて抽出した複数の構造表象を、マルチストリーム構造と呼ぶ。

マルチストリーム構造を用いると、Fig.2の処理によりマルチストリーム正規化二乗誤差行列が計算できる。これに対し、3段階重回帰分析を用いて外国語発音評価を行う枠組みをFig.4に示す。

1段階目の重回帰では、各正規化二乗誤差行列の行ごとに重回帰を行う。これは2段階重回帰の1段階目の重回帰と同様の処理である。

2段階目の重回帰では、1段階目の重回帰の結果に対し、

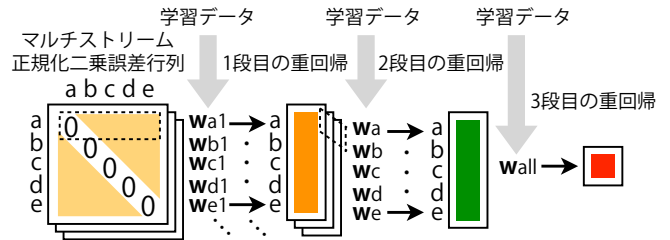


Fig. 4 3段階の重回帰を用いた外国語発音評価

各音響イベントごとに、ストリーム方向に重回帰分析を行う。重回帰分析の目的変数としては、例えば1段階目と同じ各音素に対する手動評価値が利用できる。重回帰のパラメータの学習時には、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構造を重要視するかが学習される。例えば、16kHz サンプリング音声のMFCCと6kHz サンプリング音声のMFCCを使ってマルチストリーム構造化した場合、低周波数領域に音素の特徴が多く含まれる母音は6kHz 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段階目の重回帰の結果、各音響イベントごとの評価値が得られる。

3段階目の重回帰では、2段階目の重回帰で計算した各音響イベントの評価値に対して重回帰分析を行う。これは2段階重回帰の2段階目の重回帰と同様の処理である。

2段階の重回帰と3段階の重回帰を比較すると、マルチストリーム化を行い、2段階目の重回帰によってストリームごとに重みを付けているところが差分であり、この処理により、精度がさらに向上すると考えられる。

## 4 実験

構造表象と多段階重回帰を用いた外国語発音評価の効果を実験により検証する。

実験には、ERJデータベースを用いる[11]。ERJでは、日本人大学生が約75文からなる米語読み上げ文セットを1セット読み上げている。文セットは8つあり、それぞれのセットを25名程度が読み上げている。これを、それぞれの学生ごとにTable1の条件で構造表象化した。なお、マルチストリーム構造化を行う場合には、予備実験によりMFCC(12次元)、 $\Delta MFCC$ (12次元)、16bit/6kHzでリサンプリングした音声のMFCC(12次元)の三種類の音響特徴量を利用することとした。16bit/6kHzでリサンプリングした音声を利用した理由は、母音を特徴づけるフォルマントはおよそ3kHz以下の周波数領域にあらわれるという音声学的知見に基づくものである。

Table 1 構造抽出条件

sampling	16bit / 16kHz
窓	窓長 25msec, シフト長 10msec
特徴量	MFCC(12 次元)
HMM	1 混合 monophone (対角共分散行列)
トポロジー	left-to-right, 3 状態

教師の音声には、ERJに含まれる20名の米語ネイティブ話者のうち、男性であるM08氏1名分のみを利用する。教師の構造表象を作るための特定話者音素HMMの学習には、生徒が読み上げた文章セットと同じ文章セットを利用した。

重回帰分析の目的変数には、ERJに含まれる、各学生に対し音声学者5名が採点した手動評価値の平均値を利用する。これを各学生に対する手動スコアとして、最終段階の重回帰分析の目的変数として利用する。最終段階以前の段階における重回帰分析の目的変数には、音素ごとの手動評価値があればそれを利用すべきである。しかし、ERJには音素ごとの手動評価値は付けられていないため、前段階の重回帰分析の目的変数にも、最終段階と同様、各学生に対する手動スコアを用いた。なお、重回帰分析の学習データには、全8セットのうち7セット分を用い、残りの1セットを評価データとした。

leave-1set-outで、提案手法を用いた構造表象による学生のスコアとERJに含まれる手動評価値との相関値を算出したときの平均と標準偏差を、Fig.5に示す。比較のために、[5]で提案されている部分構造分析を用いて同様の実験を行った結果も示している。また、GOPを用いた場合のスコア相関値も示している。GOPを用いて生徒のスコアを計算する際には、各音素ごとに算出されるGOPスコアに対し、多段階重回帰の最終段階の重回帰と同様の重回帰を行うことにより、生徒のスコアを算出した。なお、GOPの計算に用いるHMMの学習には、ERJに含まれるすべての米語ネイティブ話者20名の音声すべてを用い、音響特徴量はMFCC.E.D.Z.Nを用いた。さらに、構造表象とGOPの単純な組み合わせによる評価も行った。具体的には、3段階重回帰における2段階の重回帰の説明変数として、GOPスコアを追加することによりで二つの手法を組み合わせた。この結果もFig.5に示している。

結果、提案手法である多段階重回帰は、従来手法である部分構造を用いた手法と比較して高い精度が得られた(有意差あり)。2段階重回帰と3段階重回帰では、3段階重回帰の方がやや高い精度が得られた(有意差なし)。また、GOPスコアを利用した手法と提案手法を比較すると、提案手法の方がやや高い精

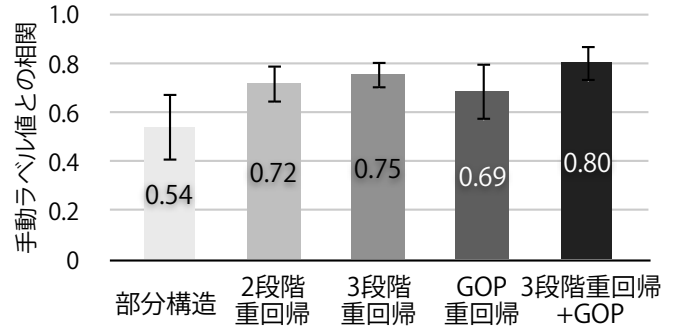


Fig. 5 米語発音自動評価値と手動ラベル値との相関

度が得られた(有意差なし)。さらに、GOPスコアと提案手法を組み合わせることで、GOPと比較して高い精度が得られた(有意差あり)。構造表象を用いた音声分析では、GOPスコアで利用されるような音の絶対的な特徴を捨て、静的な変形に対する頑健性の強い音と音との相対関係の情報のみを利用している。そのため、話者の違いによる変形が大きい有声音に関しては、構造表象を用いた手法は有効である。一方、話者の違いによる変形の小さい無声音に関しては、GOPスコアのように音の絶対的な特徴をそのまま利用することが有効になる。そのため、GOPスコアと提案手法を多段階重回帰を用いて組み合わせることで、より高い精度が得られたと考えられる。

## 5 まとめ

本論文では、構造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、従来の構造表象の部分構造を用いた外国語発音評価手法に対し、大幅に精度が向上することがわかった。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る簡単な実験結果も示した。

## 参考文献

- [1] S.Furui, Proc. ASRU, pp.1-10 (2009)
- [2] 羅他, 信学技報, SP2009-32, pp.51-56 (2009)
- [3] M. Russell *et al.*, Proc. SLATE, CD-ROM (2007)
- [4] N.Minematsu, Proc. ICASSP, pp.585-588 (2004)
- [5] M. Suzuki *et al.*, Proc. ASRU, pp.574-579 (2009)
- [6] 鎌田他, 信学技報, SP2007-36, pp.73-78 (2007)
- [7] Y. Qiao *et al.*, Proc. ASRU, pp.576-581 (2007)
- [8] 朝川他, 信学技報, SP2008-113, pp.203-208(2008)
- [9] Witt *et al.*, Speech Communication, vol.30, no.2-3, pp.95-108 (2000)
- [10] Y. Qiao *et al.*, Proc. INTERSPEECH, pp.1349-1452 (2008)
- [11] 峯松他, 日本教育工学会論文誌, vol.27, no.3, pp.259-272 (2004)