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分析

鈴木 雅之† 喬 宇‡ 峯松 信明‡ 広瀬 啓吉‡
東京大学大学院工学系研究科† 東京大学大学院情報理工学系研究科‡

1 はじめに

近年，Nintendo DSやiPhone，SNSアプリ等において，語学学習ソフトが人気を集めている．文科省が小学校での外国語活動を必修化させるなど語学学習熱が高まる中，計算機を利用した語学学習ソフトのニーズは今後ますます高まると予想される．

しかし，計算機を利用した外国語発音分析においては，発音の違いも話者の違いも音声のスペクトル包絡を変形させるため，学習データとユーザで話者性が大きく異なる場合に正しい発音分析が行えなくなるという問題がある．これは，ユーザが小学生など子供の場合に特に大きな問題になる．

近年，この問題を解決する，話者の年齢や性別におよそ不変な音声の構造的表象が提案された[1]．さらに，これを用いた外国語発音分析への応用も行われている[2, 3]．本論文では，音声の構造的表象を用いた外国語発音分析を高精度化させる手法を提案する．実験の結果，提案手法は，従来手法と比較して大幅に高精度な分析が行えることがわかった．

2 音声の構造表象

話者の違いは，ケプストラム空間における可逆な空間写像で近似できる．例えばMLLR適応は，この空間写像を線形変換と仮定することで実現される．二つの空間が可逆な空間写像で結びつけられる場合，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[4]．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．

$$f_{\mathrm{div}}(p_i, p_j) = \int p_j(\boldsymbol{x})\, g\left(\frac{p_i(\boldsymbol{x})}{p_j(\boldsymbol{x})}\right) d\boldsymbol{x} \qquad (1)$$

図1に，MFCC空間において f-divergence が不変になる様子を示す．音響イベント分布間の f-divergence は，話者の違いなどの静的な変形に近似的に不変になる．我々は，すべての音響イベント分布間の f-divergence を計算することによって得られる話者に不変な構造を，音声の構造的表象と呼んでいる．

Pronunciation analysis based on speech structure and multilayer regression
†Masayuki Suzuki ‡Qiao Yu ‡Nobuaki Minematsu ‡Keikichi Hirose †Graduate School of Engineering, The Univ. of Tokyo ‡Graduate School of Information Science and Technology, The Univ. of Tokyo

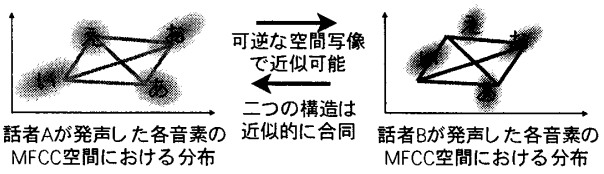

図 1: 静的な変形に不変な音声の構造的表象

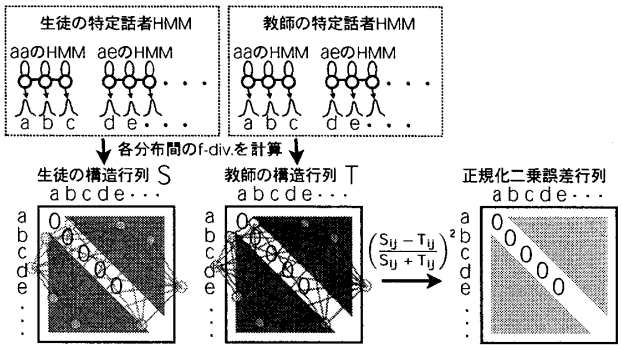

図 2: 音声の構造的表象を用いた外国語発音分析

2.1 音声の構造的表象を用いた外国語発音分析

構造表象を用いて外国語発音を分析するには，生徒の音声，教師の音声からそれぞれ構造を抽出し，生徒の構造と教師の構造の構造間にどのような差異があるか調べればよい．

図2に，生徒と教師の特定話者音素HMMから構造をそれぞれ作成し，構造間の差異を正規化二乗誤差行列によって表現する一連の流れを示す[2]．ここで，正規化二乗誤差行列とは，生徒と教師の発音がどう違うのかの情報を含む行列で，行列の各要素は

$$D_{ij}(\boldsymbol{S}, \boldsymbol{T}) = \left(\frac{S_{ij} - T_{ij}}{S_{ij} + T_{ij}}\right)^2 \qquad (2)$$

で計算される．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 f-divergence である．正規化二乗誤差行列を利用することで，生徒の発音と教師の発音の，どの部分がどの程度違うのかを定量的に分析することができる[3]．

2.2 多段階重回帰を用いた外国語発音評価

提案手法は，複数の音響特徴量を利用して構造表象をマルチストリーム化し，複数の正規化二乗誤差行列から生徒のスコアを3段階の重回帰により算出する手

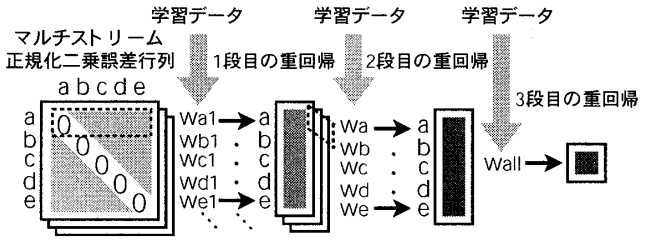

図 3: 3 段階の重回帰を用いた外国語発音分析

表 1: 構造抽出条件

sampling	16bit / 16kHz
窓	窓長 25msec，シフト長 10msec
特徴量	MFCC，ΔMFCC， 16bit/6kHz sampling 音声の MFCC
HMM	1 混合 monophone（対角共分散行列）
トポロジー	left-to-right，3 状態

法である．

まずマルチストリーム化について述べる．構造表象は，例えば MFCC 空間で f-divergence を計算し抽出されるが，例えば ΔMFCC 空間で f-divergence を計算しても抽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．複数の構造それぞれに対し図 2 の処理を行うことで，マルチストリーム正規化二乗誤差行列が抽出できる．

これに対し，3 段階重回帰分析を用いて外国語発音分析を行う枠組みを図 3 に示す．1 段目の重回帰では，各正規化二乗誤差行列の行ごとに重回帰を行う．重回帰分析の目的変数としては，例えば各音素に対する手動評価値が利用できる．2 段目の重回帰では，1 段目の重回帰の結果に対し，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 段目と同じ各音素に対する手動評価値が利用できる．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も利用できる．3 段目の重回帰では，2 段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．3 段目の重回帰の結果，生徒のスコアが得られる．

3 実験

実験には，ERJ データベースを用いる [5]．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．これを，それぞれ

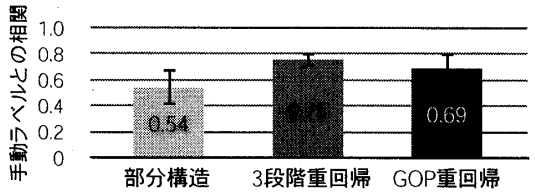

図 4: 米語発音自動評価値と手動ラベル値との相関

の学生ごとに表 1 の条件で構造表象化した．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 名分のみを利用した．重回帰分析の目的変数及び提案手法の精度評価には，学生の読み上げ音声に対し音声学者 5 名が採点した手動評定値の平均点を利用した．重回帰分析の学習には，全 8 セットのうち 7 セット分を用い，残りの 1 セットを評価に利用した．

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，図 4 に示す．比較のために，[2] 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 [6]．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．

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．また，GOP スコアを利用した手法と提案手法を比較しても，提案手法の方がやや高い精度が得られた．

4 まとめ

本論文では，構造表象を用いた外国語発音分析における精度の高い分析法として，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．

参考文献

[1] N.Minematsu, Proc. ICASSP, pp.585–588 (2004)
[2] M. Suzuki *et al.*, Proc. ASRU, pp.574–579 (2009)
[3] 鎌田他, 信学技報, SP2007-36, pp.73–78 (2007)
[4] Y. Qiao *et al.*, Proc. INTERSPEECH, pp.1349–1452 (2008)
[5] 峯松他, 日本教育工学会論文誌, vol.27, no.3, pp.259–272 (2004)
[6] Witt *et al.*, Speech Communication, vol.30, no.2–3, pp.95–108 (2000)