発音構造を用いた話者の違いに頑健な発音評定と学習者分類*

◎鈴木 雅之, 羅 徳安, 峯松 信明, 広瀬 啓吉 (東大)

1 はじめに

文部科学省は，2011 年度より小学 5・6 年生に対し，英語の「話す／聞く」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，英語を専門としない教師たちが英語の授業を担当することになる．このような状況から，子供や大人など様々な声質に対して頑健に動作する CALL システムのニーズは今後高まっていくと考えられる．

しかし，子供を対象とした発音評定には，大きな問題がある [1]．ある音素が正しく発声できているか否かと話者の声道形状の違いは，共に，音声のスペクトル包絡を変形させる．そのため，スペクトル包絡を直接利用して教師と学習者を比較すると，両者における声道形状のミスマッチに依存して，評価結果が変わってしまう．子供の声道形状は成人の声道形状と大きく異なるために，子供と成人教師のスペクトル包絡を直接比較して評価することはできない．

音声認識を目的とした場合，話者適応を行なうことでこのミスマッチを一部吸収できるが，話者適応技術を発音評定システムに直接導入すると，発音に関する適応がかかってしまい，不適切な発音を正しいと判定することが起こってしまう [2]．

筆者らはこの問題を，音声の構造的表象を利用することで解決を図っている．音声の構造的表象は，音韻群の相対関係のみにより規定されており，話者の違いに近似的に不変な性質を持つ．既に，これを用いた自動発音評定や発音誤り検出に関する検討を行なってきた [3, 4]．また，最近は音声認識への応用も検討され，構造を用いた分析手法は高度化されつつある [5]．

本研究では，一連の音声認識研究で培われた知見を生かし，発音評定手法の改善を行う．また，構造を用いない，各音韻の音響特性をそのまま用いる手法との比較実験も行う．

2 音声の構造的表象

ある二つの分布に任意の一対一対応変換を施しても，その分布間の f-divergence は常に一定となる [6]．f-divergence が不変となる概念図を，Fig. 1 に示す．Fig. 1 において，任意の写像 h に対して，$p_i(x)$ と $p_j(x)$ 間の f-divergence は $q_i(x)$ と $q_j(x)$ 間のそれと等しくなる．これは各分布の広がりの様子に応じて適切に分布中心間距離を計算することで得られる性質である．

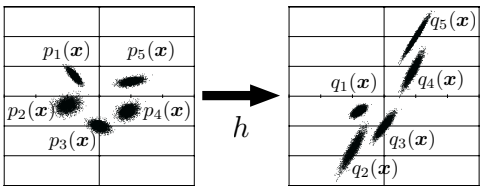

Fig. 1 変換をかけても不変な距離関係

音声の構造的表象とは，ある話者の発声を，音素などの音響イベント単位でケプストラム空間上で分布群化し，それらすべての分布間の f-divergence を計算することで得られる距離行列（幾何学的には構造）のことである．ケプストラム空間上で，話者の違いはある一対一対応変換で近似できるが，f-divergence はこの変換に対して不変であり，音声の構造的表象は話者の違いに近似的に不変となる．なお本研究では，f-divergence として Bhatacharrya Distance（BD）を利用する．

構造を用いて発音評価を行なうためには，学習者の構造と教師の構造間の差異を計算してやればよい．二つの構造間の差異は，以下の式で計算する（物理的な意味は [3] を参照）．

$$D_1(\boldsymbol{S},\boldsymbol{T}) = \sqrt{\frac{1}{M}\sum_{i<j}(S_{ij}-T_{ij})^2} \qquad (1)$$

ここで，$\boldsymbol{S}$，$\boldsymbol{T}$ は，全イベント群から計算される BD の距離行列であり，M はイベント数である．

3 提案する改善手法

本研究では，従来の構造を用いた発音評価手法 [3] に対し，三つの改善手法を提案する．

一つ目は，音響イベントの単位をより細かくすることである．従来は，音響イベント単位として音素を利用していたが，音声認識応用では，一音素あたり 3 つ程度の分布を利用する方が高い精度が得られている [5]．これを踏まえ本研究では，音素単位ではなく，音素 HMM の状態単位で距離行列を構成する．

二つ目は，次元数の削減である．構造を用いた音声認識では，PCA，LDA，ランダム選択などの方法で

*Robust estimation of proficiency and classification of learners based on pronunciation structures. by M.Suzuki, D.Luo, N.Minematsu, and K.Hirose (The University of Tokyo)

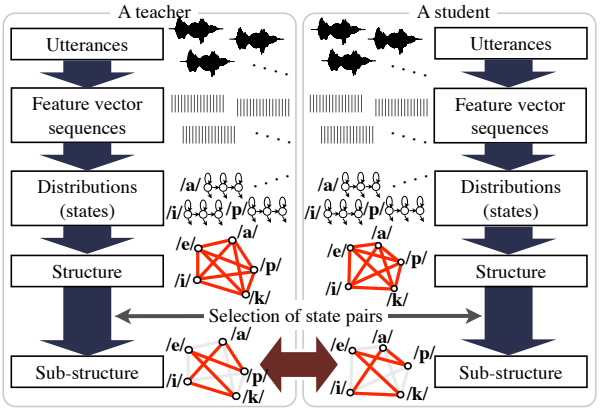

Fig. 2　学習者と教師の発声からの構造抽出と比較

次元削減を行なうことで，より高い識別能力を実現している [5]．本研究では次元削減の方法として，自動評定値と手動評定値との相関がより高くなるようにHMM 状態対（エッジ）を一本ずつ選ぶ方法を提案する．こうすることで，発音の善し悪しをより明確化するエッジを優先的に選択し，不要なエッジを省くことができる．本手法は，構造から部分構造を抽出していることになるため，以降これを部分構造化と呼ぶ．

三つ目は，構造間差異計算における相対尺度の利用である．構造のエッジには，話者平均的に長いもの，短いものが混在している．そのため，D_1（式 (1)）を使う場合，平均的に長いエッジの差異が結果に大きく反映され，短いエッジの差が無視される危険性がある．この問題を避けるため，下式を導入する．

$$D_2(\boldsymbol{S},\boldsymbol{T}) = \sqrt{\frac{1}{M}\sum_{i<j}\left\{\frac{S_{ij}-T_{ij}}{\frac{1}{2}(S_{ij}+T_{ij})}\right\}^2}. \quad (2)$$

D_2 を使うことで，エッジの相対的な差によって構造間差異を計算できる．

Fig. 2 に，提案手法を導入して発音評定する方法をまとめる．ただし，図が細かくなるのを避けるため，Fig. 2 では，構造のノードに音素を対応させているが，提案手法では，HMM の各状態をノードとする．

4　実験

4.1　データベース

実験には，ERJ データベースを用いる [7]．ERJ では，1 セット約 75 文からなる 8 つの読み上げ文セットが定義されている．200 名の日本人大学生がいずれか 1 セットを読み上げており，さらに，各話者の 10 文分に対し，音声学者 5 名が採点した手動評定値が付けられている．また，米語を母語とする教師 20 名による読み上げ音声も含まれている．本研究では M08 氏 1 名分の音声を教師の音声として使用する．

Table 1　音響分析条件

サンプリング	16bit / 16kHz
窓	25 msec 幅，10 msec シフト
学習データ	一名につき約 75 文
特徴量	MFCC（12 次元）
HMMs	不特定話者・文脈非依存モノフォン HMM
出力確率分布	対角共分散行列を持つ単一ガウス分布
トポロジー	3 状態の left to right 型

4.2　構造・GOP を用いた自動発音評定実験

音響分析条件を Table 1 に示す．200 名の学習者から，各々43 個ずつの音素 HMM を作成し，構造を抽出した．音響イベント単位として音素を選んだ場合は ${}_{43}C_2 = 903$ 本，HMM 状態を選んだ場合は ${}_{43\times3}C_2 = 8,256$ 本のエッジからなる構造である．また，学習者と比較する教師 M08 の音声からは，学習者・教師間で同じ文セットを用いて構造を抽出した．部分構造化（エッジ選択）の貪欲探索は，8 つの文セットのうち，第 6 セットを読み上げた学習者以外の音声を学習データとして用いた．自動評定値としては，学習者の部分構造と教師の部分構造間の D_1 や D_2 を計算し，符号を反転したものを用いた．手動評定値には，10 文分の評定値の平均を用いた．最後に，オープンデータである第 6 セットを読み上げた 26 名の音声データを用い，自動評定値と手動評定値との相関を計算することで，性能を評価する．

比較のため，Goodness Of Pronunciation (GOP) スコア を用いた自動評定も行なった．GOP は，Witt らによって提案された，広く自動発音評定に利用されている発音評定尺度である [8]．GOP は，ある文の読み上げ音声を観測したときの，意図された音素列に対する事後確率値で定義される．これは，以下の式で近似できる．

$$\begin{aligned} &GOP(o_1,...,o_T,p_1,...,p_N) \\ &= P(p_1,...,p_N|o_1,...,o_T) \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\max_{q\in Q}P(o^{p_i}|q)}\right\} \end{aligned} \quad (3)$$

ここで，T は観測系列長であり，N は音素数である．また，o^{p_i} は強制アライメントによって得られる p_i に対応する系列であり，D_{p_i} はその継続長である．Q は考慮している全音素種類である．音響モデルには，構造抽出と同様のデータ，すなわち教師 M08 が発声した 1 セット分の音声から学習した HMM と，ERJ に含まれる 20 名の教師の全発声を使って学習した HMM を 2 通り用いた．また音響特徴量には，MFCC に Δ 特徴量及び，Δ パワーを加えたものを用いた．

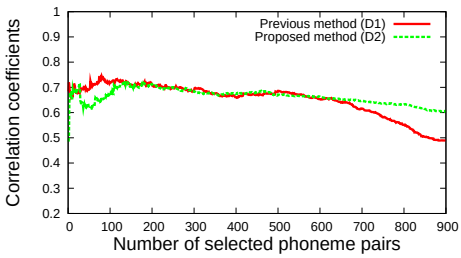

Fig. 3　音素ベースの実験結果

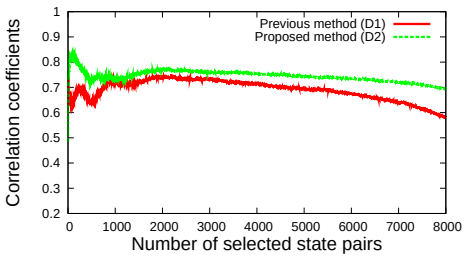

Fig. 4　状態ベースの実験結果

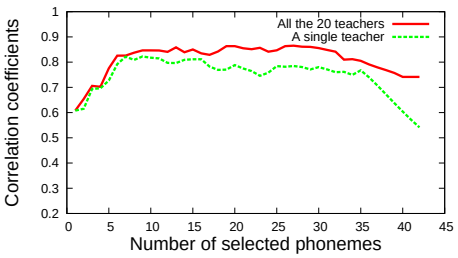

Fig. 5　GOP スコアによる結果

Table 2　選ばれたエッジの組み合わせ

L_2:R_3/ AE_2:ER_2/ L_2:R_2/ Z_2:CH_3/ Z_1:JH_2/
EY_2:OY_2/ R_3:Z_2/ L_3:AO_2/ AO_3:AW_1/ AE_3:AY_2/
Z_1:UH_3/ AX_2:IH_1/ CH_1:AXR_2/ AE_3:AX_2/ AW_1:ER_1/
G_3:ZH_2/ K_2:sil_1/ ER_2:AXR_1/ Z_2:ZH_3/ AE_3:EH_2/
W_2:DH_1/ EY_2:JH_3/ AW_3:OW_2/ DH_3:sil_2/ CH_3:EH_2/
L_2:AW_3/ CH_1:EY_2/ F_1:K_3/ S_1:OW_2/ M_3:AO_2/
AO_2:ER_2/ K_3:HH_3/ JH_3:AXR_1/ P_3:JH_2/ EH_2:UW_1/
F_2:SH_3/ D_2:sil_2/ D_3:SH_1/ R_3:sil_1/ AO_1:AW_1/
T_3:Z_1/ L_2:NG_3/ M_1:UH_3/ K_3:CH_1/ AY_1:UW_2/
L_2:L_3/ EH_2:UW_2/ L_1:IH_2/ Z_1:Z_2/ B_2:UH_2/ W_3:Z_3/
L_3:UW_1/ N_2:JH_1/ L_1:AA_3/ N_1:V_1/ AH_1:IH_3/
B_1:HH_2/ L_1:Y_1/ S_1:sil_3/ S_2:V_1/ AE_2:EY_3/
AX_2:EY_2/ AW_2:AX_2/ CH_1:IY_2/ D_1:CH_2/ F_1:S_1/
AH_3:OW_3/ K_3:JH_2/ AA_2:NG_3/ J_3:EH_3/ UH_1:AXR_2/
AW_1:HH_3/ M_2:Y_2/ L_1:N_3/ AX_2:EY_3/ AE_3:IH_1/
AA_1:UW_1/ R_3:AO_3/ K_3:DH_3/ R_3:UH_3/ N_1:IH_3/
F_1:EH_3/ AE_2:IH_1/ AX_2:EH_2/ G_1:ZH_2/ DH_3:OW_1

4.3　実験結果

音素をノードとする構造を用いた結果を Fig. 3 に，状態をノードとする構造を用いた結果を Fig. 4 に示す．横軸は選択エッジ数を表す．色の違いは，構造間比較手法の違いであり，赤線は構造間差異に従来から用いられてきた D_1 を使った結果，緑線はエッジ間の相対的差異を用いた D_2 を使った結果である．

Fig. 3・Fig. 4 に共通して，全体的に左上がりの傾向がある．特徴量選択によって相関が向上していることが分かる．特に，状態単位の構造で D_2 を用いた場合に，部分構造化の効果は特に大きい．両図を比較すると，状態単位で構造を作成した Fig. 4 の方がより良い結果を示していることも分かる．D_1 と D_2 の違いは，提案手法である状態単位での構造化においては，D_2 を用いる方が精度が向上している．結局，状態単位で構造抽出，86 本のエッジ選択，比較尺度に D_2 を利用した場合に，相関値 0.84 で最大となった．なお，ここで部分構造として選ばれた 86 本のエッジからなる部分構造のノードには，43 音素中 41 音素が含まれている．選ばれたエッジを，選ばれた順に Table 2 に示す．日本人の苦手とする/L/と/R/，/Z/と/JH/，/ER/や/AX/や/AXR/の発音などが，上位に選ばれていることがわかる．

Fig. 5 に，GOP を用いて発音評定を用いた場合の結果を示す．赤線は ERJ に含まれる教師の全音声を用いて学習した HMM を用いた結果，緑線は構造抽出と同じデータを用いて学習した HMM を用いた結果である（教師 M08）．GOP を使った実験でも，構造の特徴量選択と同じ要領で，音素選択を行なった．横軸は選択音素数である．

結果，学習データに全音声データを用いた方が相関が高くなった．また，GOP でも音素選択が有効に働いていることも分かる．最終的に，学習データに全教師 20 名の全発声を用いた HMM を利用し，音素を 27 個選んだときに，相関値が最大で 0.87 となり，構造を用いた場合よりもやや高い相関値が得られた．

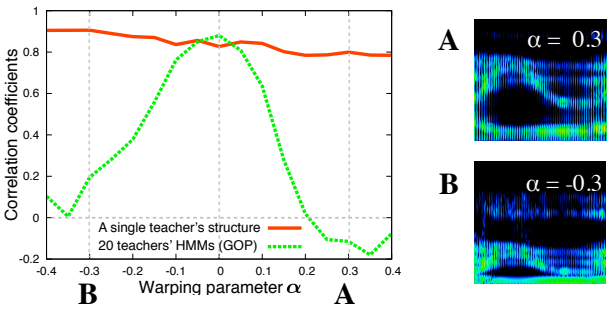

Fig. 6　声道長を変化させたときの相関値の変化

4.4　声道長ミスマッチ条件における頑健性

Fig. 6 に，人工的に声道長変換をかけた音声を評定した場合の相関値の変化を示す．赤線は，構造を用いた先の実験で最も良い性能を示した，状態単位の構造，エッジを 86 本選択，D_2 を利用した場合の結果である．緑線は，GOP を用いた先の実験で最も良い性能を示した，全教師の全音声を使った HMM を使い，音素を 27 個選んだ場合の結果である．声道長変換は，STRAIGHT を用いて周波数ウォーピングを施すことで実現した．図の横軸のウォーピングパラメータ α は，声道長変換の度合いを示すパラメータであり，α=+0.40 ／ −0.40 の時に，声道長が約半分／倍になることに対応している．

構造は，声道長の変化に対し極めて高い頑健性がある．一方 GOP は，事後確率であるため一般的には学習データと評価データの声道形状ミスマッチをキャンセルする効果があると考えられているが，大きく相関値が低下している．これは，GOP 算出時に参照する強制アライメントの精度が落ちることが原因と考えられる．

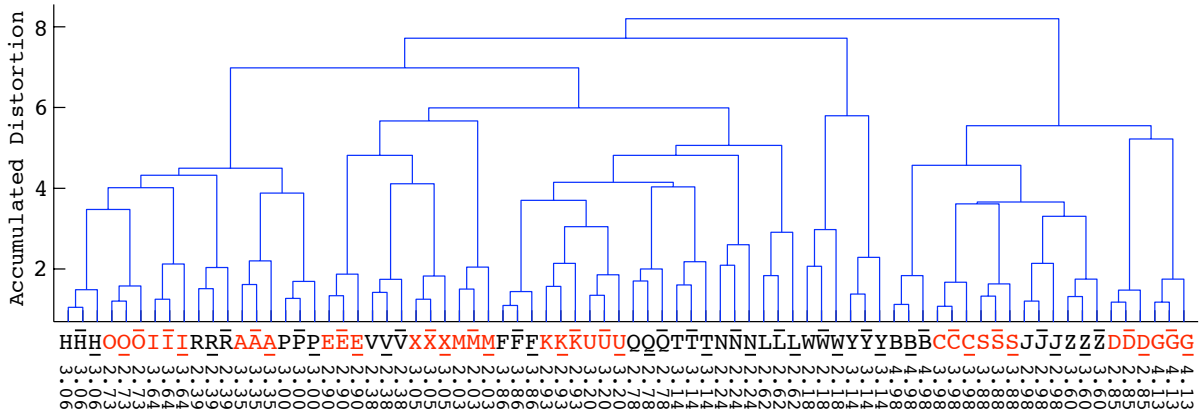

Fig. 7　構造間差異（D_2）を用いた 26 × 3 名の学習者の分類

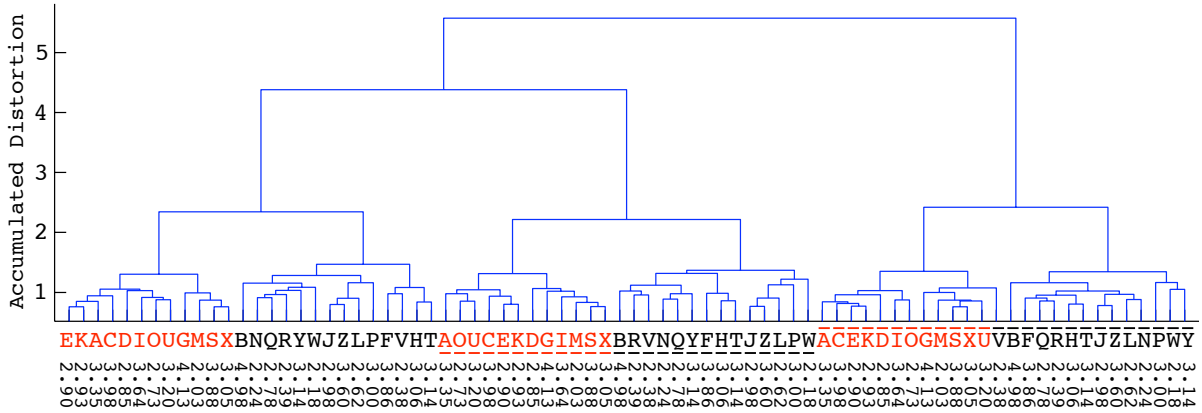

Fig. 8　スペクトル包絡の差異（D_3）を用いた 26 × 3 名の学習者の分類

4.5　学習者の分類

これまで，学習者と教師の構造間の差異により発音評定を行ってきたが，学習者間の構造間差異を使えば，学習者を分類することも可能である．先の実験で利用してきた部分構造間差異を利用して ward 法によるクラスタリングを行った結果を Fig. 7 に示す．比較のため，式 (4) を用いて，スペクトル包絡の直接の差異に基づくクラスタリングの結果を Fig. 8 に示す．

$$D_3(\boldsymbol{S},\boldsymbol{T}) = \sqrt{\frac{1}{M}\sum_i \mathrm{BD}(s_i^S, s_i^T)} \qquad (4)$$

なお，s_i^T は，話者 T の HMM の状態 i の分布を表す．分類する話者には，セット 6 に含まれる学習者 26 名すべてと，それらに対しウォーピングパラメータ $\alpha = \pm 0.30$ をかけたもの計 $26 \times 3 = 72$ 名の学習者を用いた．アルファベットが学習者の ID，赤が女性，黒が男性であり，下線付が $\alpha = +0.30$（小人），上線が $\alpha = -0.30$（巨人）を表す．数字は，手動評価値の平均を表す．

結果，絶対量を用いた分類では，性別差，ウォーピングパラメータの差といった，声道長の差によって学習者が分類されていることがわかる．一方構造を用いた分類では，声道長にはほとんど影響を受けず，話者の発音状態のみに依る分類ができている．

5　結論

本研究では，音声の構造的表象を用いた自動発音評定の改善を行なった．1) より細かな音響イベントの利用，2) 部分構造化，3) 構造比較における相対尺度を利用することで，従来手法より精度を向上させることができた．さらに，音響的特性を直接用いる手法と比較して，話者の違いに頑健な処理を実現できることを示した．

参考文献

[1] M. Russell *et al.*,*Proc. SLaTE*, CD-ROM, 2007.

[2] 羅 他，信学技報，SP2009-32，pp.51–56，2009.

[3] 峯松，信学技報，SP2003-180，pp.31–36，2004.

[4] 朝川 他，電子情報通信学会論文誌，vol.J90-D，no.5，pp.1249–1262，2007.

[5] Y. Qiao *et al.*,*Proc. ASRU*, pp.576–581, 2007.

[6] Y. Qiao *et al.*, *Proc. INTERSPEECH*, pp.1349–1452, 2008.

[7] 峯松 他，日本教育工学会論文誌，vol.27，no.3，pp.259-272，2004

[8] S. M. Witt *et al.*, Speech Communication, 30, pp.95–108, 2000.