Vol.2009-SLP-77 No.17
2009/7/18

音声の構造的表象を用いた自動発音評定法の改善

鈴木雅之[†1]　羅徳安[†1]
峯松信明[†1]　広瀬啓吉[†1]

外国語の発音学習において，意図した音素が音響的に適切に生成されているか否かは，多くの場合スペクトル包絡を通して推定可能である．しかし，学習者の性別，年齢などの個人差もスペクトル包絡を変形させるため，これらの要因が発音評定精度を低下させることがある．近年，音声に含まれる言語的特徴と非言語的特徴を分離することを目的として，音声の構造的表象が提案され，発音評定において高い有効性が示されている．この提案後，我々は構造的表象を音声認識に対しても応用し，分析の高度化を図ってきた．本研究では，これまでに得られた知見を元に，構造を用いた発音自動評定の改善を検討した．さらに，構造比較における相対的尺度を新しく導入した．実験の結果，手動評定値と自動評定値の相関が従来手法より高くなること，GOP などの既存の手法と比較して，個人差に対して高い頑健性を示すことが分かった．

Improvement of Structure-based Automatic Estimation of Pronunciation Proficiency

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

Adequacy in controlling the vocal organs is often estimated from spectral envelopes of input utterances but the envelope patterns are also affected by alternating speakers. To develop a good and stable method for automatic estimation of pronunciation proficiency, the envelope changes caused by linguistic factors and those by extra-linguistic factors should be properly separated. For this aim, a structural representation of pronunciation was proposed recently and its effectiveness was experimentally shown. After the proposal, we have tested that representation also for ASR and, through these works, we have learned better how to apply speech structures to various tasks. In this paper, based on our recently acquired knowledge on the structures, several methods are examined to improve the automatic estimation of pronunciation proficiency. Further, a relative structural distance measure is also proposed. Experimental results show that higher correlations are obtained between human rating and machine rating and that, in comparison to widely-used GOP scores, higher robustness is realized with respect to extra-linguistic factors.

1. はじめに

近年，CALL システムが広く用いられるようになってきている．特に，Nintendo DS や iPhone をはじめとする携帯端末上で動作する英語学習ソフトウェアは，高い人気を集めている．この背景には，携帯端末の普及に加え，日本人自身の英語能力への不安感，そして英語教育への期待の高まりがある．文部科学省は，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 年生の英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．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，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ますます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，大きな問題があることが知られている[1)]．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，(成人の音声資料より構築したシステムを用いた場合）音声認識率や発音評定精度の低下を招き易い．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，MFCC のようなスペクトル包絡を表現する特徴量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，比較結果が変わることがある．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る[2)]．これは，発音の善し悪しも話者の個人差もスペクトル包絡という同一の物理量によって表象されるからである．

結局，音声に含まれる言語的側面と非言語的側面を分離する方法論が求められることになるが，近年，これを実現する音声の構造的表象が提案された[3)]．音声の言語的側面のみを話者性に頑健に表象する訳であるが，この性質は，声の個人性をその個人のデフォルト声道形状がもたらす音色のバイアス項として捉え，発声（音色）のダイナミクスを，変換不変量を

†1 東京大学
The University of Tokyo

Vol.2009-SLP-77 No.17
2009/7/18

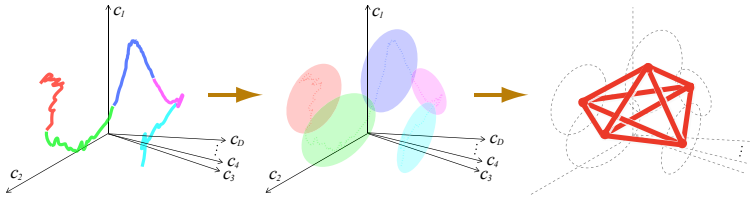

図 1 f-divergence によって作られる一発声の構造的表象
Fig. 1 An utterance structure composed only of f-divergences

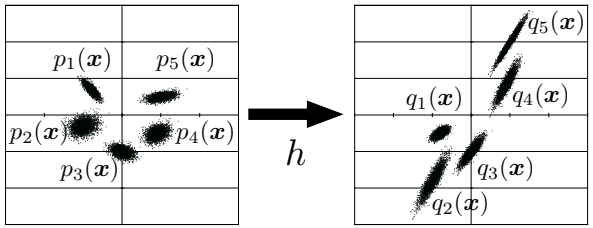

図 2 変換をかけても不変な距離関係
Fig. 2 Speaker-invariant system of language sounds

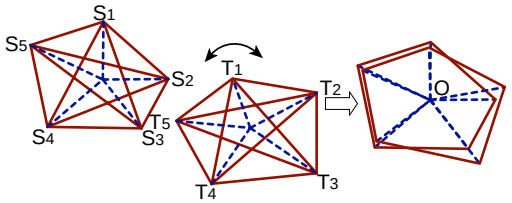

図 3 二つの構造の比較
Fig. 3 Structure comparison through shift & rotation

用いて表象することで得られる．すなわち，音声中の音響イベントの絶対的音響量を捨象し，イベント群から成る距離行列を用いて，発声（イベント群）を構造として表象する．これを用いて外国語発音を表象すると，個人差の大部分が消失し，音韻の幾何学構造のみが浮き彫りになる．既に，自動発音評定や発音誤り検出に関する検討を行なって来た[3),4)]．最近で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),6)]．

本研究では，音声の構造的表象を用いた自動発音評定を取り扱う．具体的には，峯松が 2004 年に行なった実験を再度試みる[3)]．先行研究との差分は，構造に基づく音声認識研究の中で得られた種々の知見を取り入れ，更なる精度向上を図ったことである．1) 音素より細かな音響イベント単位の利用，2) 特徴量選択による部分構造化を検討し，さらに，3) 二つの構造間差異を相対的に計算する手法を新たに導入する．

2. 音声の構造的表象を用いた分析

音声の構造的表象を一発声から抽出する方法を**図 1** に示す．まず一発声からケプストラム時系列を抽出し，それを自動区分化し，各区分を分布としてモデル化することで，音響イベント分布群を得る．そして，それらの音響イベント間の f-divergence（分布間距離尺度の一種）を計算することで，一つの幾何学構造を定義する．図 1 は，一発声からの構造抽出を図示しているが，複数発声からの構造抽出も可能である．例えば複数の発声から，特定話者音素 HMM を学習し，各音素 HMM の出力確率分布群を音響イベント群として構造を抽出する方法がある[3)]．他には，英語の単母音を含む単語を発声させ，各母音部分を切り出して分布化したものを音響イベントとして，構造を抽出することも可能である[7)]．

次に，f-divergence の性質について述べる．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[8)]．f-divergence が不変となる概念図を，**図 2** に示す．図 2 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空間を局所的に歪めて分布中心間距離を計測することで得られる性質である．本研究では，f-divergence（の関数）として，Bhattacharyya Distance（BD）の平方根を使用している．二つの正規分布 $\mathcal{N}_a(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)$，$\mathcal{N}_b(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b)$ 間の BD は，下記となる．

$$\mathrm{BD}\left(\mathcal{N}_a, \mathcal{N}_b\right) = \frac{1}{8}\left(\boldsymbol{\mu_a} - \boldsymbol{\mu_b}\right)^T \left(\frac{\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b}}{2}\right)^{-1} \left(\boldsymbol{\mu_a} - \boldsymbol{\mu_b}\right) + \frac{1}{2}\log\frac{\left|\left(\boldsymbol{\Sigma_a} + \boldsymbol{\Sigma_b}\right)/2\right|}{\left|\boldsymbol{\Sigma_a}\right|^{\frac{1}{2}}\left|\boldsymbol{\Sigma_b}\right|^{\frac{1}{2}}} \quad (1)$$

構造を用いて音声分析を行なうためには，二つの構造間を比較する尺度が必要になる．ケプストラム空間において，マイク特性差異と声道長差異は，およそケプストラム軌跡に対するシフト・回転という幾何学的変換に対応することになる[9)]．このことを踏まえ，二つの構造を比較する概念図を，**図 3** に示す．二つの構造間の距離は，最も値が小さくなるように適切にシフト・回転を行なった後の，全ての頂点間の距離の和として定義する．これは，以下の式で非常によく近似できることが実験的に示されている[3)]．

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

ここで，$\boldsymbol{S}$ と $\boldsymbol{T}$ は，全イベント群から計算される f-divergence の距離行列であり，M はイベント数である．式 (2) を利用することで，構造の回転やシフト（すなわち適応処理）を明示的に行なわずに，適切な回転・シフト後のスコアが得られることになる．

以上の手法を用い，学習者構造と教師構造の比較を通して，学習者習熟度の自動評定が可能になる．既に，構造による自動評定値と，English Read by Japanese database（ERJ）[10)] に含まれる手動評定値間の，高い相関関係が確認されている[3)]．さらに $D_1(\boldsymbol{S}, \boldsymbol{T})$ を各音響イベントペアに分解することで，矯正対象音素を特定する手法も提案されている[4),7)]．

Vol.2009-SLP-77 No.17
2009/7/18

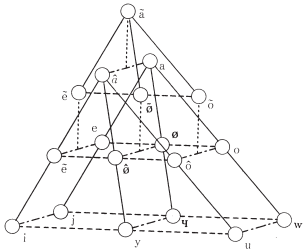

図 4 ヤコブソンによる仏語母音体系
Fig. 4 The French vowel system proposed by R. Jakobson

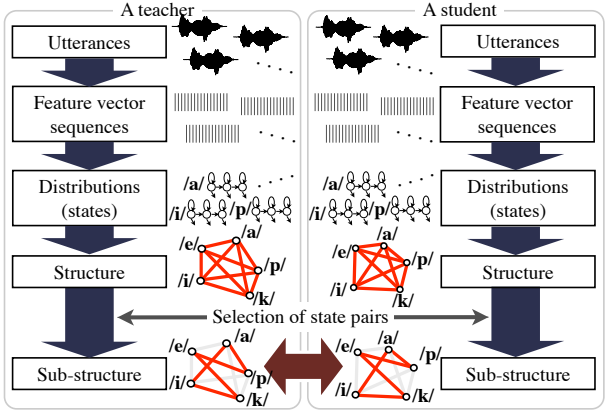

図 5 学習者と教師の音声からの部分構造抽出と比較
Fig. 5 Sub-structure extraction for a student and a teacher

3. 発音の構造的表象の言語学的解釈

ここで，発音の構造的表象の言語学的解釈について示しておく．学習者発声を構成する一音一音に対して，適切な音響的特性（スペクトル包絡）が観測されるか否かを検討する場合，これは（音響）音声学的な評定手法と解釈できる．一方提案手法は，各音のスペクトル包絡ではなく，個々の音が他の音群と適切な関係に結ばれているか否かを評定する方法である．言語学的に考えれば，これは構造主義的音韻論[11] に基づく評定手法と解釈出来る．外国語発音の習得過程を，個々の音を学習するのではなく，適切な音の体系を学習する過程として捉える．個々の音をそのまま学習する（模倣する）過程は声帯模写であり，これは発音学習とは質的に異なる音声活動である．**図 4** にヤコブソンによる仏語母音群の体系を示す．

4. 音声の構造的表象を用いた自動発音評定の改善手法

本研究では，音声の構造的表象を用いた自動発音評定を取り扱い，その分析手法を改善することで精度の向上を図る．本節では，従来用いられてきた構造を用いた自動発音評定手法に対し，具体的にどのような点で改善を行なうのかについて述べる．

一つ目は，音響イベント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採択していた．具体的には，3 状態の音素 HMM を学習し，音素間距離を三つの分布（状態）間距離の平均として定義し，音素距離行列を導出していた．しかし，構造を用いた音声認識では，HMM 同様，一音素あたり 3 つ程度の分布を用意することでより高い精度が得られている[5),6)]．これを踏まえ自動発音評定においても，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ことで，精度向上を検討する．

二つ目は，次元数の削減である．構造を表現する特徴量の次元数，幾何学的には構造のエッジ（対角線）数は，音響イベント数 M に対して $M(M-1)/2$ となり，M^2 のオーダーで増加する．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)]．本研究においても，適切な次元削減を行なうことでより高い精度が得られる可能性がある．ここ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エッジ選択する方法を採用する．具体的には，自動評定値と手動評定値との相関を評価関数として，エッジを一つずつ貪欲に選択することを繰り返す．こうすることで，発音の上手／下手をより明確化するエッジが優先的に選択され，不要なエッジは省かれることになる．本手法は，幾何学的には，構造から部分構造を抽出していることになるので，以降，この処理を部分構造化と呼ぶ．

以上の改善手法に加えて，構造間差異計算における相対的尺度の利用を提案する．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，式 (2) 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．

$$D_2(\boldsymbol{S},\boldsymbol{T}) = \sqrt{\frac{1}{M}\sum_{i<j}\left\{\frac{S_{ij}-T_{ij}}{\frac{1}{2}(S_{ij}+T_{ij})}\right\}^2}. \tag{3}$$

式 (3) 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する．

図 5 に，本研究の提案手法を導入して教師・学習者の音声から各々構造を抽出し，比較する方法をまとめる．まず，複数の読み上げ音声から，特定話者音素 HMM を学習する（音響イベント分布を作成できるのであれば他の処理に変更可能）．次に，f-divergence の距離行列を計算することで，構造を得る．図が細かくなるのを避けるため，図 5 では音素を 5 つのみとし，さらに構造のノードに音素を対応させる従来手法を示しているが，提案手法では，構造ノードは音素 HMM の各状態に対応する．次に，適切なエッジを選択することで部分構造を作成する．以上のプロセスを教師と学習者各々に行い，最後に二つの部分構造間差異 D_2 を計算することで，学習者と学習者の発音を幾何学的に比較する．従来の CALL システムでは母語話者の不特定話者音響モデルが用いられて来たが，提案手法では特定話者の教師構造を使用する．これは学習者が好みの英語教師を選択できることを意味している[12)]．

表 1 音響分析条件
Table 1 Conditions for acoustic analysis

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
音素の種類	aa,ae,ah,ao,aw,ax,axr,ay,b,ch,d,dh,eh,er,ey,f,g,hh,ih, iy,j,jh,k,l,m,n,ng,ow,oy,p,r,s,sh,t,th,uh,uw,v,w,y,z,zh,sil 合計 43 種類

5. 実　　験

5.1 データベース

実験には，ERJ データベースを用いる[10]．ERJ では，8 つの読み上げ文セットが定義されている．各文セットは，TIMIT に含まれる文，日本人にとって難しい発音が含まれる文など，約 75 文によって構成されている．ERJ には，いずれか 1 セットに対する，200 名の日本人大学生の読み上げ音声が収録されている．各学習者の 10 文発声に対して，日本人学習者の癖をよく理解している，米語を母語とする音声学者 5 名が採点した手動評定値も含まれている．また，これらの音声データと手動評定ラベルに加え，20 名の米語を母語とする教師による読み上げ文音声も含まれている．なお，8 セット全てを読み上げているのは，20 名中 2 人である（M08&F12）．本研究では M08 を教師音声として使用している．

5.2 構造・GOP を用いた自動発音評定実験

音響分析条件を**表 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，文セット毎に音素 HMM を構築し，構造抽出を行なった．これは，学習者・教師間で同一文セットによる構造比較を行なうためである．結局，学習者 200 名から抽出した 200 の構造と教師から抽出した 8 の構造，計 208 の構造を抽出した．

部分構造化（エッジ選択）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．エッジ選択は，貪欲探索を用い，自動評定値と手動評定値との相関が高くなるように一つずつエッジを追加していくことで行なった．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．手動評定値は各学習者が持つ 5 名の音声学者による 10 文分の評定値の平均を，その学習者の手動評定値とした．こうして定義された部分構造に対して，オープンデータである第 6 セットを読み上げた 26 名の音声データを評価データとして用い，自動評定値と手動評定値との相関を求めた．なお，第 6 セットを評価データにした理由は，他セットと比べて，第 6 セットの学習者の手動評定値が，最も幅広く分布していたためである．

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 らによってで提案された，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[13]．GOP は，ある文の読み上げ音声を観測した時の，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．

$$\begin{aligned} &GOP(o_1,...,o_T,p_1,...,p_N) = P(p_1,...,p_N|o_1,...,o_T) \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\sum_{q\in Q}P(o^{p_i}|q)}\right\} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{D_{p_i}}\log\left\{\frac{P(o^{p_i}|p_i)}{\max_{q\in Q}P(o^{p_i}|q)}\right\} \end{aligned} \quad (4)$$

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．ここで，$\{o^{p_1},...,o^{p_N}\}$ は $\{o_1,...,o_T\}$ に対応している．Q は考慮している全音素種類である．GOP は事後確率値で定義されているため，教師 HMM と学習者の音声でミスマッチがあったとしても，およそキャンセルできるスコア計算となっている．本研究では，GOP 算出のための音響モデルとして，9 種類の HMM を用意した．うち 8 つは，構造抽出と同様のデータ，すなわち教師 M08 が発声した，8 つの文セットのうち 1 つから学習した HMM である．もう一つは，ERJ に含まれる 20 名の教師の全発声を使って学習した HMM である．また音響特徴量としては，MFCC に Δ 特徴量及び，Δ パワーを加えた 25 次元の特徴量を使用した．

5.3 実 験 結 果

音素単位で構造を構成した結果を **図 6** 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を意味している．赤線は，構造間差異に従来から用いられてきた式 D_1 を使った結果であり，緑線は，エッジ間の相対的差異を用いた式 D_2 を使った結果である．

次に提案手法である，状態単位で構造を構成した結果を**図 7** に示す．横軸は，先と同様，選択エッジ数である．色の対応も同じである．

図 6・図 7 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図 7 の方がより良い結果を示し

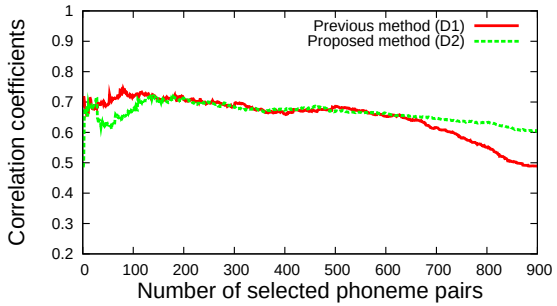

図 6 音素ベースの構造分析による相関値
Fig. 6 Correlations with phoneme-based structure analysis

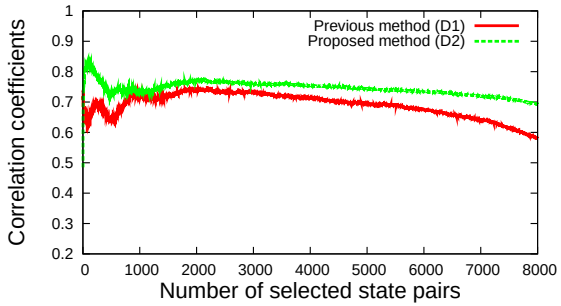

図 7 状態ベースの構造分析による相関値
Fig. 7 Correlations with state-based structure analysis

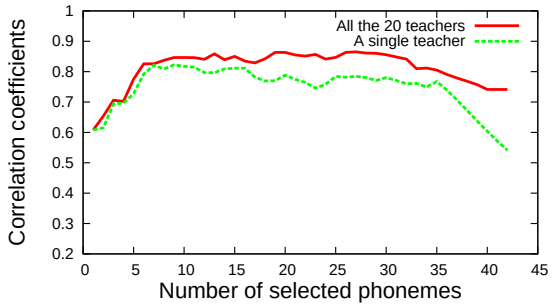

図 8 GOP スコアによる相関値
Fig. 8 Correlations with GOP analysis

ていることも分かる．D_1 と D_2 の効果であるが，図 6 と図 7 で傾向が異なる．しかし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 本のエッジをみると，全 43 音素のうち，41 音素に関係したエッジが抽出されていた．

手動評定を行なった 5 名の音声学者の評定値に対し，ある 1 名の評定値と残り 4 名の評定値の平均値の相関係数を計算すると，相関が高い順に 0.94，0.92，0.91，0.87，0.83 となっており，提案手法は手動評定値と同等の精度・安定性があるといえる．

図 8 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．

実験の結果，学習データに全音声データを用いた方が相関が高くなることが分かる．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．

5.4 声道長ミスマッチ条件における頑健性

図 9 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

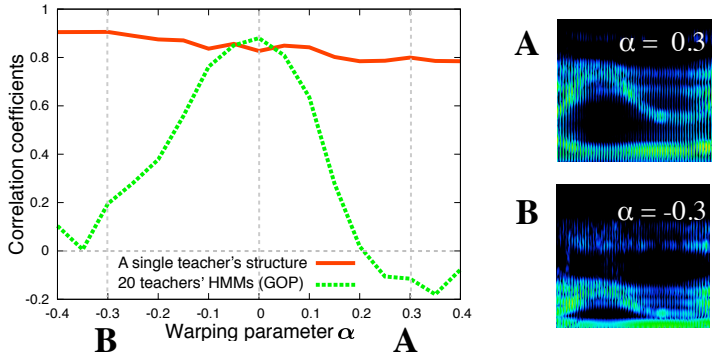

図 9 ウォーピングされた音声を利用した場合の相関値
Fig. 9 Correlations with warped utterances

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．また，α=+0.30 ／ −0.30 時のスペクトログラム例を図示している．周波数ウォーピングが，スペクトログラムを大きく変化させていることが分かる．このような変化にも拘らず，構造を用いたスコアは極めて高い頑健性を見せている．一方，GOP を用いた場合は，大きく相関値が低下しており，や

がて相関値はゼロになる．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るが，実際には，GOP の頑健性は非常に低い．これは，ミスマッチが置きた場合，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である．これを防ぐためには，教師 HMM を学習者音声に話者適応することになるが，第 1 節に示したように，これは別の問題を引き起こす．市販されている小学生用の CALL システムは学習言語を母語とする小学生の音声データベースを用いて構築されている[14]．技術的観点から考察すると，(成長期にあり声変わりもする小学生高学年の）ある学習者にとって，最も適切な教師 HMM は，その学習者が対象言語を習得した時に発声される音声データから学習される HMM となる．

6. 考　　察

音声認識をタスクとした場合，ある話者に対して最も精度が高い音響モデルは，その話者の大量の音声データによって学習された音響モデルである．発音評定をタスクとした場合でも上記の技術的要請から明らかなように，GOP などのスコアを用いる場合，最も精度が高くなる教師音響モデルは，当該学習者の音声から学習される音響モデルとなる．音声認識にしろ，発音評定にしろ,「音そのもの」に対する音響尤度を用いるシステムを構築すれば，話者が変われば声の音響特性が変わる以上，上記は常に成立する事実である．

結局，音響尤度が示しているのは，ある音と別のある音（あるいは別のある音分布）との距離であり，これを用いて発音評定すれば，教師の発声を如何に物理的に真似できたかの評定することになる．これは第 3 節でも述べた様に，発音の習熟度推定ではなく，本来は，声帯模写の技能獲得度を推定するために使われるべき枠組みである．

男性教師の「Repeat after me」に対して，どうにか太い声を出そうとする小学生は通常いない．声帯模写を通して外国語を学ぶ例は通常見られない．にもかかわらず，従来の CALL システム開発の多くは，技術的には，外国語発音能力の習得過程を，声帯模写技能の習得過程と同一視しているかのような開発が行なわれている．我々の提案する CALL システムは,「教師の発声によって伝達された音素列情報を，自らの調音器官を使って生成する場合に，教師音声の何を真似るべきなのか」という問いに対する物理的な回答を準備した上で開発を行なっている．

2011 年以降，日本国籍を持つ全ての小学生 5,6 年生が「話す／聞く」英語教育を，英語教育とは無縁であった教師から，受けることになる．何らかの技術支援を考える場合，科学的妥当性，技術的妥当性の両方を備えた枠組みを導入すべきであると考える．

7. 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精度の改善を行なった．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，その声道長差異に対する頑健性を実験的に確かめた．特に，GOP と比較して構造を用いた自動発音評定は，声道長差異に対して高い頑健性を持つことを実験的に示すことができた．

参　考　文　献

1) M. Russell *et al.*, “Challenges for computer recognition of children’s speech,” *Proc. SLaTE*, CD-ROM, 2007.
2) 羅徳安 他，“シャドーイング・音読発音評価を目的とした話者適応の分析と応用”，信学技報，SP2009 (2009-6，発表予定)
3) 峯松信明，“音声の音響的普遍構造の歪みに着眼した外国語発音の自動評定”，信学技報，SP2003-180，pp.31-36 (2004-1)
4) 朝川智 他，“音声の構造的表象に基づく英語学習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 (2007-5)
5) Y. Qiao *et al.,* “Random discriminant structure analysis for continous Japanese vowel recognition,” Proc. ASRU, pp.576–581, 2007.
6) S. Asakawa *et al.,* “Multi-stream parameterization for structural speech recognition,” *Proc. ICASSP*, pp.4097–4100, 2008.
7) N. Minematsu *et al.,* “Structural representation of the pronunciation and its use for classifying Japanese learners of English,” *Proc. SLaTE*, CD-ROM, 2007.
8) Y. Qiao *et al.,* “f-divergence is a generalized invariant measure between distributions,” *Proc. INTERSPEECH*, pp.1349–1452, 2008.
9) D. Saito *et al.*, “Directional dependency of cepstrum on vocal tract length,” *Proc. ICASSP*, pp.4485–4488, 2008.
10) N. Minematsu, *et al.*, “Development of English speech database read by Japanese to support CALL research,” *Proc. ICA*, pp.577–560, 2004.
11) ローマン・ヤコブソン他，“言語音形論”，岩波書店 (1986)
12) 高澤真章 他，“音声の構造的表象に基づく発音評価とその応用”，音講論，3-10-12，pp.489–492 (2008-3)
13) S. M. Witt *et al.*, “Phone-level pronunciation scoring and assessment for interactive language learning,” Speech Communication, 30, pp.95–108, 2000.
14) ベネッセ小学生向け英語学習プログラム BE-GO
http://be-go.benesse.ne.jp/be-go/