調音特徴に基づく音素単位での英語発音誤り検出と発音評価

長岡　紘昭†　入部　百合絵‡　桂田　浩一†　新田　恒雄†
豊橋技術科学大学　大学院工学研究科†
豊橋技術科学大学　情報メディア基盤センター‡

1. はじめに

英語の発音学習は，通常，会話学校において対面学習により学ぶか，発音に関する書籍やCDをもとに学習するのが一般的である．しかし，英会話学校では費用と拘束時間の問題が，また書籍やCDによる自習では，誤りを的確かつ具体的に指摘し，同時に矯正方法を指導する評価者の不在という問題がある．一方，音声認識技術を利用した英語発音学習の研究も数多く提案されている[1][2]．提案された研究の多くでは，特徴量にMFCCを採用するため，多量の音声コーパスを必要としていた．他方，調音特徴は話者共通の調音運動を表現するため，特徴抽出に多くの音声データを必要とするが，音素や単語列推定には，少量の学習データで高精度音声認識を実現できる[3]．本報告では，特徴量に調音特徴（Articulatory Feature; AF）を採用することでMFCCよりも高精度な音声認識を目指す．加えて，学習者の音声から調音特徴を抽出することで，発音誤りを音素毎に指摘可能な英語発音学習システムを開発する．学習者の調音運動を元に改善方法を提示できると，自学自習で効率よく発声を矯正することができる．

2. システムの構成

英語発音学習システムは，調音特徴抽出部，音素認識部，発音評価部から構成される．以下，各処理について述べる．

2.1. 調音特徴抽出部

調音特徴は，発音学習での利用を考慮し，IPA（国際音声記号）から英語に関する部分を抽出して作成した．調音特徴は，音声スペクトル系列の時間微分と周波数微分から求めた局所特徴（Local Feature; LF）を多層ニューラルネット（Multi-Layer Neural network; MLN）に入力して得る．MLN学習では，時刻t-3，t，t+3の3フレームを入力し，音素環境の違いに対応した調音特徴を教師信号として学習する．認識段階には，出力の調音特徴系列から，調音の方法と調音部位の情報が得られるため，学習者の発音とその矯正点を容易に把握できる．

2.2. 音素認識部

調音特徴をHMMに通し，音素列と音素境界を得る．発音学習では，予め正解音素列と共に，誤り易い音素列の組み合わせをネットワーク文法の形で構成しておき，これらを参照しながらHMMで音素列と尤度を計算している．学習者が誤り易い音素列としては，以下に示す日本人が誤りやすい五つの例を考慮し構成している．

(1) 有声/無声，母音/長母音の区別
　有声音と無声音，母音と長母音の置き換え
　例）bag /b ae g/ を [b ae k]（バック）と発音
(2) 子音置換
　英語の子音のうち，日本語で対応する音が存在しないもの
　例）sea /s iy/ を [sh iy]（シー）と発音
(3) 母音置換
　英語には複数のア（aa, ax, ah, ae）が存在するが，日本語では1種類の音（ア）しか存在しないなど（実際には発音しているがそれらを区別していない）
　例）map /m ae p/ を [m ah p]（マップ）と発音
(4) 母音挿入
　子音閉鎖音や子音連続時に母音を挿入
　例）get /g eh t/ を [g eh t ao]
　　（ゲット）と発音
(5) rの脱落
　音節末の /r/ が脱落
　例）far /f er/ を [f aa]（ファー）と発音

以上の規則を誤り音素列とした場合，例えば単語read /r iy d/ における正解音素列と誤り音素列の組み合わせは，16通りになる（図1参照）．

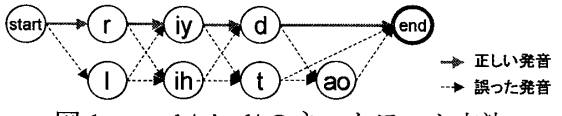


図1　read /r iy d/ のネットワーク文法

Phoneme recognition and English pronuciation evaluation based on Articulatory Feature
†Graduate School of Engineering, Toyohashi Univ. of Tech.
‡Information and Media Center, Toyohashi Univ. of Tech

2.3. 発音評価部

HMM が出力する音素列を参照し，発音の正確さを音素毎に調音特徴のスコアから評価する．発音評価値 S の算出式を以下に示す．

$$S = \left(d_{corr}(f) - d_{mis}(f)\right) \ / \ \max(d(f))$$

ここで，f は HMM が出力する音素境界の中心フレームを示している．また，$d_{corr}(f)$は正解音素にのみ存在する調音の素性(正解音素 r と l の例では，接近音)のスコア，$d_{mis}(f)$は誤り音素にのみ存在する調音の素性(r と l の例では，側面接近音)のスコアを示している．即ち，正解音素の調音特徴と誤り音素の調音特徴を比較し，各々固有に存在する調音特徴のスコアの差を，最大の調音スコア max(d(f))で正規化することにより，正解あるいは誤りの度合いを示す．

具体的には，学習者が単語 read （/r iy d/）を，[l iy d] と誤って発音した場合，発音評価値（この場合は，発音の誤り度）は（0.6-0.1）/ 0.9 ≒ 0.56 となる(図 2 参照)．

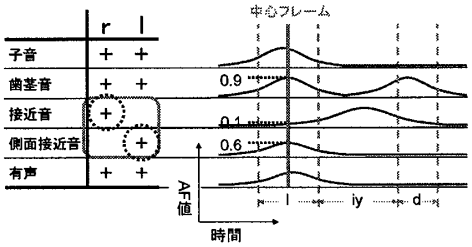
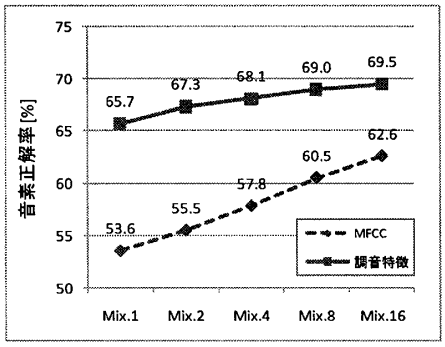


図 2　スコアの計算例

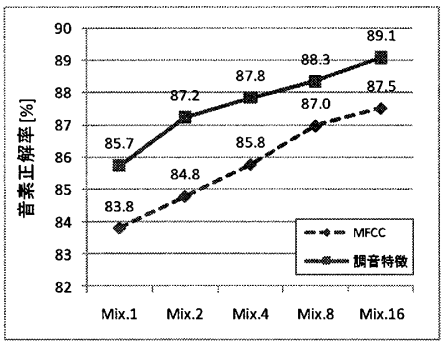
3.　評価実験

本報告では，特徴量に調音特徴（84 次元）を用いたため，一般的な特徴量である MFCC(Δ，ΔΔ，ΔP，ΔΔP；38 次元)との比較実験を行った．また，ネットワーク文法により出力音素列を拘束しているため，拘束有り/無しの英語音素正解率を比較した．MLN の学習には，TIMIT [4]の 4,240 文（男性話者 376 名，女性話者 154 名）を使用した．音響モデルは，モノフォン（英語 42 音素），5 状態，3 ループで，混合数を 1，2，4，8，16 と変化させて実験した．音響モデルの学習には，MLN 学習と同じ TIMIT のデータを使用した．テストデータは TIMIT の 800 文（未知の男性話者 62 名，女性話者 38 名）である．

実験結果を図 3 に示す．拘束なしでは（図 3(a)），調音特徴の方が MFCC よりも約 7%～12%高い正解率を得た．音素列拘束ありの場合（図 3(b)）においても，調音特徴の方が MFCC よりも約 2%高い正解率を得ている．音素列の拘束有り/無しにかかわらず，調音特徴の方が優位な結果となった．また，音素列拘束ありの場合は拘束なしよりも約 20 %以上正解率が向上している．特徴量に調音特徴を採用し出力音素列を拘束したことで，高い音素正解率を得ることができた．今回の実験に使用した音声データ TIMIT の話者は，全て英語を母語とするアメリカ人である．今回の実験から，英語母語話者に対しては高い確率で音素を認識できることがわかった．しかし，音素列の拘束に使用したネットワーク文法は，日本人の英語発音誤りを反映したものであり，今後，日本人の英語発音コーパスを使用した評価実験を行い，音素列拘束をした場合に発音誤りを正しく抽出しているか検証する必要がある．



(a) 音素列拘束なし



(b) 音素列拘束あり

図 3　調音特徴と MFCC の比較

4.　まとめ

調音特徴に基づく音素認識を用いた英語発音学習システムを提案した．調音特徴と MFCC について，音素正解率の比較を行った結果，ネットワーク文法による出力音素列拘束有り/無しにかかわらず調音特徴が優位な結果を示した．今後は，英語教師による発音評価とシステムが出力した発音評価との比較実験を行いたい．

参考文献

[1] Hongcui Wang, Speech Communication, Vol.51, No.10, pp.995-1005, 2009.
[2] 河合，他，日本音響学会誌，Vol.57，No.9，pp.569-580, 2001
[3] 新田，他，情処技術報告 SLP，77 (4)，pp.1-6，2009.
[4] J.S.Garofolo et al., Linguistic Data Consortium, 1993.