Vol.2010-SLP-80 No.13
2010/2/13

英語発音学習のための調音特徴抽出と発音評価

長岡紘昭[†]　入部百合絵[††]　桂田浩一[†]　新田恒雄[†]

CALL 教材の開発が，近年盛んに行われるようになり，英語を学ぶ日本人学生を対象とした，様々な英語発音学習システムが開発されている．本報告では，発音学習に効果的と考えられる，調音特徴ベースの音素認識エンジンとのその評価について述べる．提案システムは，音声から調音特徴を抽出した後，HMM で調音運動を表現し音素を認識する．調音特徴抽出には，2 段構成の多層ニューラルネット(NLN) を用いた．音素 HMM では，調音特徴系列で音響モデルを表現すると共に，日本人の英語発音誤り傾向をもとに，ネットワーク文法を設計して出力音素列を制約している．英語母語話者に対する評価実験を，標準的に用いられる音響特徴量の MFCC と比較して行った結果，ネットワーク文法の有無に依らず，調音特徴が優位であった．本文では，日本人学生による英語発音評価結果についても示す．

Articulatory Feature Extraction and Its Evaluation for English Pronunciation Learning

Hiroaki Nagaoka[†], Yurie Iribe[†], Kouichi Katsurada[†]
, and Tsuneo Nitta[†]

Various types of CALL systems have been developed in recent years. We have been developing an English pronunciation learning system for Japanese students. In this paper, firstly, a phoneme recognition engine in the system is described. The engine converts an input utterance into an articulatory feature (AF) sequence using an AF extractor with two-stage multi-layer neural networks (MLN), then the AF sequence is fed into phoneme-HMMs in which the phoneme strings are constrained by a network grammar designed using easily mistakable pronunciations by the Japanese. In the experiments on TIMIT corpus, AF outperforms the acoustic feature of MFCC when the evaluations are done with and without a network grammar. Experimental results on English word utterances by Japanese students are also described.

1. はじめに

英語の発音学習は，通常，会話学校で対面学習を通して学ぶか，発音に関する書籍[1] や CD をもとに学習するのが一般的である．しかし，英会話学校は費用と拘束時間の問題が，また書籍や CD による自習は，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．そこで，自主学習に使用でき，学習者の発音評定も可能な CALL 教材の開発が盛んに行われている．

市販されてた CALL 教材としては，発音力 [2] や発音美人 [3] などがある．発音力は，学習者が発声した単語に対して，一つの音素に注目して音素認識を行い，その結果を学習者にフィードバックする．この場合，一つの音素を学習するには適しているが，他の音素が間違った場合，それを検知しないという問題がある．発音美人は，発声した単語中のすべての音素に対して，発音の良否を教示してくれるが，どのように発音すれば良いかといった矯正方法は教えてくれない．我々が開発を進めている発音学習システムでは，学習者の発話中の全ての音素に対して認識を行い，結果を学習者にフィードバックする．また，学習者がどのように発音すれば良いか，といった矯正方法も教示することを目標としている．

一方，音声認識技術を利用した英語発音学習の研究も盛んに行われている [4] [5] [6]．坪田ら [7] は，日本人の英語発音誤りをネットワーク文法の形式でパターン化し，認識の際はそのパターンに合った音素列のみを対象に音素認識をすることで，音素正解率を向上させた．今回我々が開発する発音学習システムにおいても，認識の際，日本人の英語発音誤り傾向から設計した，ネットワーク文法を用いて出力音素列を制限する．

我々は，日本人学生を対象に英語発音学習システムを開発することを目的としている．一般的に，音素認識に用いられる特徴量は MFCC であるが，本システムでは調音特徴を採用している．MFCC では，十分な音素認識率を達成するために多量の音声コーパスを必要とする．他方，調音特徴は話者共通の調音運動を HMM に表現するため，特徴抽出には多量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な認識を達成できる [8]．これまで，日本語を対象とした mono-phone HMM では，調音特徴ベース HMM が MFCC ベース HMM を上回る性能を得ている [9]．本報告では，英語発音学習システムに組み込む音素認識エンジンとその評価結果を述べる．

[†] 豊橋技術科学大学 大学院工学研究科
Graduate School of Engineering, Toyohashi University of Technology
[††] 豊橋技術科学大学 情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology

Vol.2010-SLP-80 No.13
2010/2/13

2. 調音特徴に基づく音素認識

以下に調音特徴について説明した後，調音特徴抽出から音素認識までの処理を説明する．

2.1 調音特徴

調音特徴（Articulatory Feature; AF）は，単音分類に用いられる調音様式（母音，子音，有声，無声など）と調音位置（前舌，半狭，半広，など）の諸属性を指す．AF では，あらゆる音素は調音特徴の有無(+/ -) を示すベクトルで表現できる．AF を音声認識で利用する際の利点は，調音的に近い音素同士を距離の近いベクトルとして表現できることである（例えば，[p] と[b] ではスペクトル構造は大きく異なるが，AF では有声，無声の違いのみである）．

今回用いた調音特徴セットは，国際音声記号（International Phonetic Alphabet: IPA）[10] から英語に関する部分を取り出し作成した．表 1 に，使用した 28 次元の AF セットを示す（次元数 28 次元，音素数 42 (sil を含む)）．例えば，音素[p]は，IPA では両唇音でかつ破裂音であるため，該当箇所に”＋”が付く．ここに述べた調音特徴セットは，後述するニューラルネット学習で教師信号として用いられる．

2.2 局所特徴の抽出

図 1 に調音特徴抽出器の全体構成を示した．AF 抽出器に入力された音声は，まず局所特徴 [11]（Local Feature, LF）に変換される．LF の抽出手順を図 2 に示す．入力音声は，16kHz でサンプリングされた後，25ms のハミング窓で 10ms 毎に，512 点の FFT 処理を受ける。この結果はパワースペクトルの形で積分され，中心周波数を（聴覚に近似した）メル尺度間隔で設計した 24-ch の BPF (Band Pass Filter) 出力にまとめられる。ここまでが分析処理である。続いてパワースペクトル系列上の音響特徴抽出が行われる。パワースペクトル系列が構成する曲面は，多様体として見ると時間と周波数方向の局所的な微分要素で表現できる（微分多様体)。そこで，BPF 出力を 3×3 の局所特徴に変換するため，時間軸と周波数軸上で各々3 点の線形回帰 (Linear Regression; LR)演算を行い，微分特徴としての LF を抽出する。二つの局所特徴は各 24 次元であるが，続いて離散余弦変換 (Discrete Cosine Transform; DCT)処理によって半分の 12 次元に圧縮される。これに対数パワー成分の微分要素を加えた 25 次元の特徴が LF である．

2.3 調音特徴の抽出

2.3.1 局所特徴から AF への写像

LF は，1 段目の MLN（MLN_1）によって AF へ変換される．入力の LF と出力の AF には，ともに注目フレームx_t と前後 3 点離れたフレーム（x_{t-3}, x_{t+3}）を用いた．

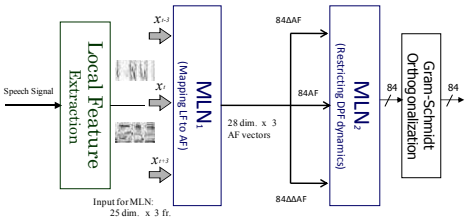

図 1　調音特徴抽出器の構成

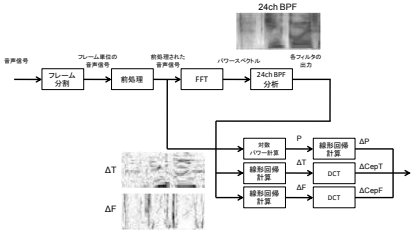

図 2　局所特徴抽出過程

表 1　調音特徴セット（左: 子音，右: 母音）

	p	b	t	d	k	g	f	v	th	dh	s	z	sh	zh	ch	jh	hh	m	n	ng	r	l	dx	w	y
母音																								+	+
子音	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
有声		+		+		+		+		+		+		+		+		+	+	+	+	+	+	+	+
無声	+		+		+		+		+		+		+		+		+								
両唇音	+	+																							
唇歯音							+	+																	
歯音									+	+															
歯茎音			+	+							+	+			+	+		+		+	+	+			
後部歯茎音													+	+	+	+									
口蓋音																									+
軟口蓋音					+	+															+		+		
声門音																	+								
破裂音	+	+	+	+	+	+									+	+									
鼻音																		+	+	+					
はじき音																								+	
摩擦音							+	+	+	+	+	+	+	+	+	+	+								
接近音																						+		+	+
側面接近音																							+		

	iy	ih	ey	eh	ae	aa	ay	aw	ao	ow	oy	uh	uw	ah	er	ax
母音	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
子音																
有声	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
前舌	+	+	+	+	+		+	+								
中舌															+	+
後舌						+		+	+	+	+	+	+	+		
狭	+	+	+				+	+		+	+	+	+			
半狭			+	+					+							+
半広				+	+					+		+		+	+	
広						+	+	+								
円唇									+				+			
r の音色																+
緊張		+	+				+	+		+	+	+				

すなわち，入力は 75 次元（25×3）の LF，出力は 84 次元（28 × 3）の AF である．注目フレームx_t だけでなく，前後 3 点離れたフレーム(x_{t-3}, x_{t+3}) のスペクトル情報を含むことで，AF への変換精度が向上する．学習はラベル付き音声データを用いて行い，＋の属性を 0.9，－の属性を 0.1 とし，誤差逆伝播法を用いた（以降，MLN_2 についても同様）．

2.3.2 調音運動による AF の整形

MLN_1 の出力を 2 段目のMLN_2 で，さらに整形する．ここでは AF の Δ（速度）と Δ Δ（加速度）の運動成分を利用して整形を行う．1 段目で抽出した AF 系列は，調音運動（= 調音器官の動き）を反映していると考え，Δ（速度）と Δ Δ（加速度）を用いて，より円滑に運動状態を表現する．なお，ここでも 1 段目同様，x_{t-3}, x_t, x_{t+3} の 3 フレームを入出力とする．すなわち MLN の入力は，AF，Δ，Δ Δ をそれぞれ 3 フレームで 252 次元，出力が AF84 次元となる．図 3 に，MLN_2の出力（AF）例を示す．発話内容は Daphne’s Swedish /d ae f sil n iy s sil w iy dx ih sh/（110 フレーム）である．

次に，ここまでに抽出された AF 系列は，各時刻間で互いに相関を持つ．対角共分散のみを使用する HMM では，無相関化が必要になるため，MLN_2の出力に対して，Gram-Schmidt の直交化を適用して無相関化を行う．

Vol.2010-SLP-80 No.13
2010/2/13

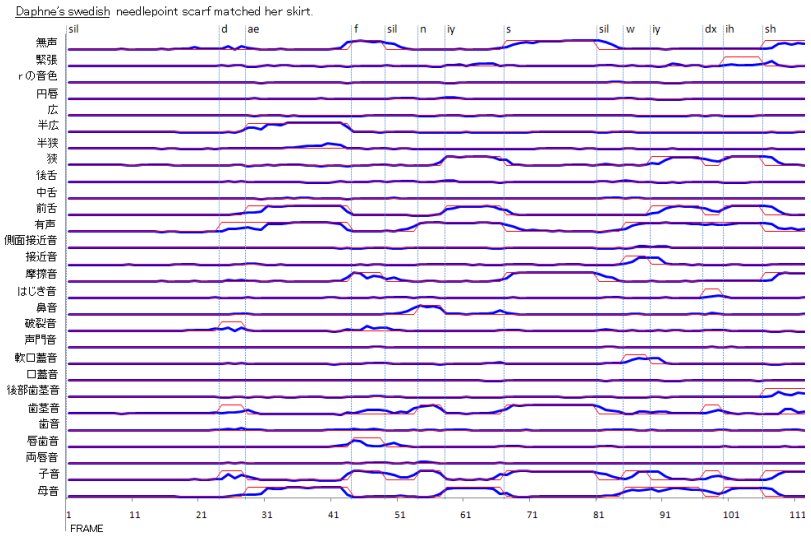

図 3　MLN_2 の出力（細線：理想値，太線：MLN_2 出力）

2.4 HMM による音素認識

直交化調音特徴系列を HMM に入力して音素列を得る．発音学習では，予めシステムが提示した単語を学習者が発音する形式を採るため，HMM から得られる音素列に対して，日本人が間違い易い発音誤りをルール化したネットワーク文法の制約を適用する．

2.4.1. 日本人の英語発音誤り

日本人の英語発音では，以下に示す発音誤りが多い．

(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音）
　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

以上の規則を適用した誤り音素列は，単語 read /r iy d/ を例にとると，正解列と誤り列の組み合わせは 16 通りになる（図 4 参照）．

2.4.2. ネットワーク文法の生成

2.4.1. で説明した規則から生成した，正解音素と誤り音素の組合せを表 2 に示す．/r/ が正解音素の場合，誤り音素は /l/ であるから，/r/ と /l/ の 2 音素になる．/ey/ が正解音素の場合，誤り音素は /eh ih/, /eh iy/ である．2.4.1.(4)で示した母音挿入時に挿入される母音は，語尾子音が /t/, /d/ の場合 /ao/，語尾子音が /ch/, /jh/ の場合 /ih/，その他の子音の場合 /uh/ である．単語の語尾子音が /d/ だとその後に挿入される母音は /ao/ になる（read /r iy d/ だと /r iy d ao/ になる）．子音連続時の子音が t だと，その後に /d/ が挿入される（handle /h ae n d l/ だと /h ae n d ao l uh/ になる）．日本人の英語発音誤り傾向をネットワーク文法に組み入れることで，HMM から出力される音素列に制約を入れた．候補を制限することで，音素数が固定され，挿入誤りが防げる．置換される単音の候補も減少するため，認識誤りが減少すると考えられる．

表 2　正解音素と誤り音素の組合せ

正解音素	誤り音素
p	b
b	p
t	d
d	t
k	g
g	k
f	v
v	f, b
th	dh, s, sh
dh	th
s	z, sh
z	s, zh, jh
sh	zh
zh	sh
ch	jh
jh	ch
hh	
m	
n	m, ng
ng	n, n g
r	l
l	r
dx	
w	
y	
iy	ih
ih	iy
ey	eh + ih iy
eh	ey
ae	ax, aa, ah
aa	ah, ax, ae
ay	ah ax aa ae + ih iy
aw	ah ax aa ae + uh uw
ao	ow
ow	ao + uh uw
oy	ao + ih iy
uh	uw
uw	uh
ah	ax, aa, ae
er	ax
ax	ah, aa, ae

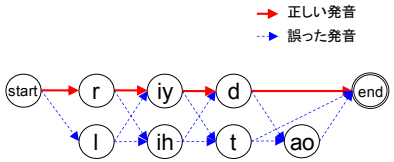

図 4　read の文法

Vol.2010-SLP-80 No.13
2010/2/13

3. 評価実験

3.1 調音特徴の評価

調音特徴の抽出精度を確認する実験を行った．評価は MLN から抽出した AF 84 (28 × 3) 次元のうち，中心フレームに相当する 28 次元に対して行う．AF 抽出精度は，次に示す特徴当たりの正解率 (AF-Correct Rate; AFCR) で計算した．

$$\mathrm{AFCR} = \frac{\text{正しく抽出できた属性数}}{\text{フレーム数} \times 28} \times 100\ [\%]$$

図 5 に AF 抽出精度を示す．MLN_1およびMLN_2のいずれの抽出過程でも 90% を越える精度が得られた．MLN の段数を増やことで精度も向上し，95%の抽出精度が得られる．MLN_2見られたの入力に運動量のΔとΔΔを含めた効果が大きいことがわかる．音素毎の AF 抽出精度を調査したところ，全ての音素においてMLN_2 の方がMLN_1 よりも高い精度であった．特に ay, aw, oy, ow などの二重母音が顕著に改善されていた．二重母音は，その他の音素と異なり二つの音を発音するため音素区間内での AF 値の変化が大きい．MLN_2は，ΔとΔΔを入力に含めているので，AF 値の変化を上手くとらえることができたと考えている．

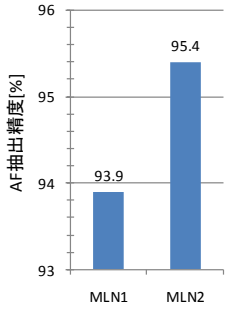

図 5　AF 抽出精度

3.2 英語母語話者に対する音素認識性能

本報告では，特徴量に調音特徴（84 次元）を用いたため，より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制約しているため，制約有り/無しの場合についても音素認識性能を調べた．音声試料は次の 2 セットを用いた．

D1: 学習セット　（MLN 学習用，音響モデル学習用）
TIMIT [13] 4,240 文，
男声 376 名，女声 154 名 (16 kHz, 16 bit)

D2: 評価セット
TIMIT 800 文，
男声 62 名，女声 38 名 (16 kHz, 16 bit)

HMM は 5 ステート 3 ループの標準的な left-to right 型を使用した．単音(mono-phone)単位で，混合数を 1, 2, 4, 8, 16 とし，D2 セットの音素認識性能を調べた．音素列制約なしの結果を図 6 に示す．調音特徴の方が MFCC よりも，Correct Rate は約 7%～12%，Accuracy は約 13～25%優位となった．調音特徴 AF は，混合数も 1 混合で高い Correct Rate を達成している．これに対して MFCC は，混合数を増やすほど向上する．この結果から，調音特徴は話者不変のパラメータであることが示唆される．

音素列制約ありの結果を図 7 に示す．Mix.16 における音素毎の Correct Rate の比較を図 8 に示す．図 8 で Correct Rate が 100%の音素は，ネットワーク文法における誤りが存在しない音素(/hh/, /m/, /dx/, /w/, /y/)である．

図 7 の Correct Rate のグラフから，調音特徴，MFCC ともに 9 割近く正しく音素を認識できることが示された．ネットワーク文法を使用しない場合よりも約 20%近く Correct Rate が上昇した．候補を制限することで，置換誤りが減少したためと考えられる．Accuracy に関しては 25%以上上昇した．候補を制限することで，音素数が固定されるため挿入誤りが大きく減少したと考える．

調音特徴の方が MFCC よりも，Correct Rate は約 2%，Accuracy は約 2%優位となった．ネットワーク文法を使用しない場合（図 6）よりも，Correct Rate, Accuracy ともに MFCC と調音特徴との差が縮まった．これは，誤り音素の組合せが少ない場合は MFCC と調音特徴に大差はなく，組み合わせが多くなる場合は，調音特徴の方が優位であることを表している．現に，図 8 では誤り音素の組合せが最も多い音素である子音（/th/, /s/, /z/, /n/）は調音特徴の方が優位であることが示されている．母音（/aa/, /ah/, /ae/. /ax/, /ay/, /ow/）に関しても，/aw/, /oy/を除き調音特徴の方が優位である．

Vol.2010-SLP-80 No.13
2010/2/13

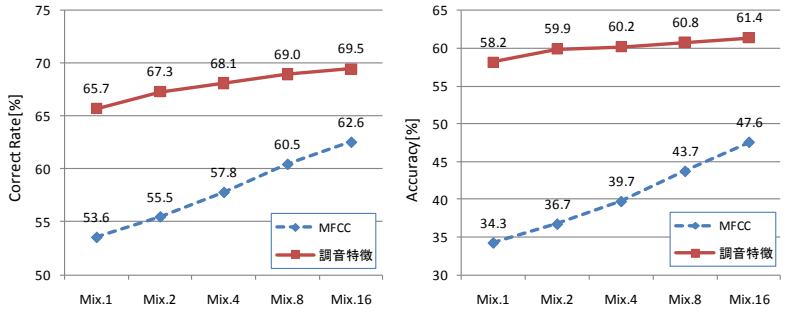

図 6　MFCC と調音特徴の音素認識結果の比較（音素列制約なし）

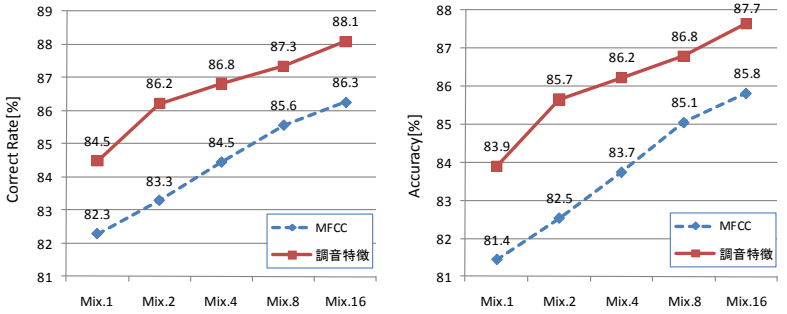

図 7　MFCC と調音特徴の音素認識結果の比較（音素列制約あり）

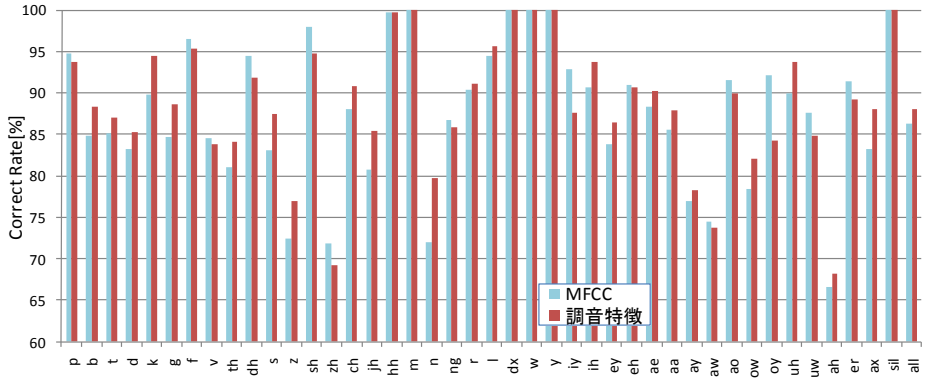

図 8　音素毎の Correct Rate（TIMIT，音素列制約あり）

3.4 日本人学生の英語音声による音素認識性能

3.2 および 3.3.の実験により英語母語話者に対しては高い確率で音素認識できることがわかった．しかし，今回音素列の制約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基に作成したものであるため，日本人が英語を発声したコーパスを使用した評価実験を行い，音素列制約をした場合に発音誤りを正しく検出できているかを検証した．

特徴量は，3.3. の実験で最も結果の良かった調音特徴(84 次元) を使用した．HMM は 5 ステート 3 ループの標準的な left-to right 型を使用した．単音(mono-phone)単位で，混合数を 16 とし，D3 セットの音素認識性能を調べた．

D3: 評価セット

日本人学生による孤立英単語発音 5988 文（話者一人当たり約 850 単語発音），男声 2 名，女声 5 名 (16 kHz, 16 bit)

学習セットは D1 を用いた．ネットワーク文法による音素列制約をした場合としない場合の比較結果を図 9 に示す．音素毎の Correct Rate を図 10 に示す．評価セット D3 には，日本人学生の発声に対し英語教師のラベリング（音素単位）が付属されている．Correct Rate は，英語教師のラベリングとシステムが出力した音素列とを比較し，算出した．図 10 の中には Correct Rate が 0%の音素（/dx/, /er/）が存在するが，これは評価セット D3 には出現しない音素である．/sil/ に関しては計測していない．

評価セットに TIMIT を使用した場合と比較し，Correct Rate は 20%低下した．音素毎にどの音素と誤って認識されているのかを調査したところ，子音の無声音（/p/）を有声音（/b/）として誤認識している場合が多かった（/t/ と /d/, /k/ と /g/, /s/ と /z/ に関しても同様）．このため /p/, /t/, /k/, /s/ は Correct Rate が低くなった．逆に，子音の有声音（/b/, /d/, /g/, /z/）は無声音として誤認識されることはほとんどなく，Correct Rate は 90%以上と高かった．MLN の学習にアメリカ人話者の音声データ（TIMIT）を使用しているため，日本人特有の発音に対しては，うまく調音特徴を抽出できなかったと考える．今後，調音特徴抽出用 MLN に日本人学生の英語音声を使用するなどの改良を検討したい．

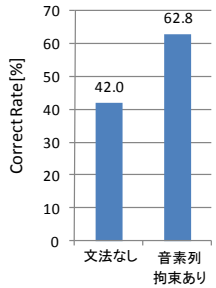

図 9 音素列制約あり/なしの Correct Rate（日本人学生の英語音声）

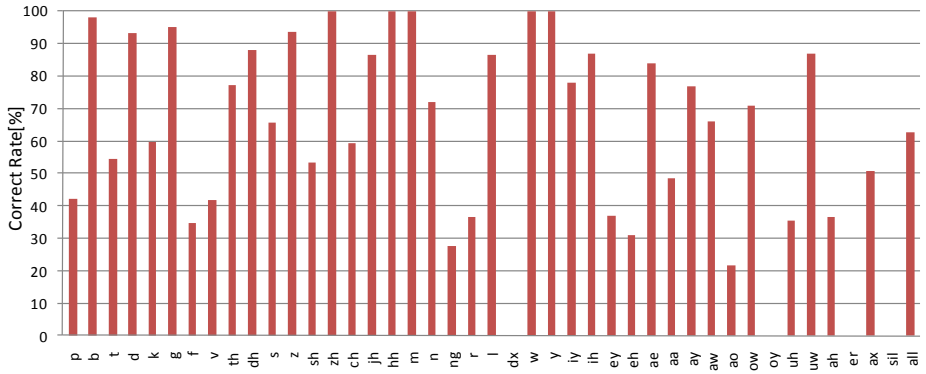

図 10 音素毎の Correct Rate（日本人学生の英語音声，音素列制約あり）

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，音素認識エンジンとしての性能を評価した．調音特徴と MFCC の両者について，音素正解率を比較した結果，ネットワーク文法による出力音素列制約の有り/無しにかかわらず，調音特徴が優位な結果を示した．一方，日本人学生の英語発話音声を使用した実験では，音素認識で良い結果は得られなかった．今後，調音特徴抽出用 MLN に日本人学生の英語音声を使用するなどの改良を検討したい．

謝辞 本研究の一部は，豊橋市新事業創出等支援事業によった．また本研究を行うにあたり，日本人学生の英語音声コーパスを提供してくださった産業技術総合研究所情報技術研究部門の児島宏明氏に感謝する．

参考文献

[1] 松澤喜好: 英語耳，アスキー（2004）
[2] 発音力）発音力／発音矯正ソフト - 株式会社プロンテスト
http://www.prontest.co.jp/soft/
[3] 発音美人/英語教材/英会話学習ソフト
http://www.wingsr.com/wing17.html
[4] 河合剛，石田朗，広瀬啓吉: 2 言語の音響モデルを用いた音声認識による非母語発音誤りの検出と発音評価，日本音響学会誌, vol.57, no.9, pp.569-580 (2001)
[5] 前田直子，山下洋一: 日本語・英語音素モデルを用いた英単語発音評定方法の検討，電子情報通信学会技術研究報告 SP, Vol.101, No.604, pp.79-86(2002).
[6] 五十里慎吾, 他: ユーザー発話のセグメンテーションと発話評価機能をもつ英語学習支援システム，情報処理学会研究報告 HI，Vol.97，No.10，pp.7-12(2002)
[7] 坪田康，壇辻正剛，河原達也: 日本人の誤りパターンの対判別を利用した英語発音教示システム，電子情報通信学会技術研究報告 SP，Vol.100，No.595，pp.25-32 (2001)
[8] 新田，他，調音運動 HMM に基づくワンモデル音声認識合成，情報処理学会研究報告 SLP，Vol.77，No.4，pp.1-6(2009)
[9] Huda, M.N., Kawashima, H. and Nitta, T., Distinctive Phonetic Feature (DPF) extraction based on MLNs and Inhibition/ Enhancement Network, IEICE Trans. Inf. & Syst., Vol.E92-D, No. 4, pp.671-680 (2009).
[10] IPA Fullchart
http://www.langsci.ucl.ac.uk/ipa/IPA_chart_(C)2005.pdf
[11] 新田恒雄，他: 複合音響特徴平面に基づく音声認識のための局所特徴抽出法，電子情報通信学会論文誌，Vol.J83-D-II, No.11, pp. 2341-2349 (2000).
[12] 福田隆，山本航，新田恒雄: 弁別的特徴ベクトルを用いた音声認識に関する検討．日本音響学会 2002 年秋季研究発表会講演論文集, Vol. I, No. 1-9-1, pp. 1-2 (2002).
[13] Garofolo , J.S. et al.: TIMIT Acoustic Phonetic Continuous Speech Corpus, Linguistic Data Consortium (1993).