F-023

日本人英語音声認識における発話者英語能力別の効果
Speaker Adaptation Performance Comparison corresponding to English Skill

松永　昭一†　　**小川　厚徳**†
Shoichi Matsunaga　　Atsunori Ogawa

1．まえがき

非母国語話者の発声は発話者の母国語発音の影響を受けるため，母国語話者の発声と比べて発声が多様となる．そのため，認識対象となる言語の母国語話者の発声で学習した音響モデルでは十分な認識性能を達成できない場合も多く，これまでも多くの性能向上の検討が行われてきた[1-3]．同様に日本人の英語が日本語の発音に影響を受けやすいという点を考慮した検討も行われている[4,5]．日本人の英語発声に対しては，日本語音素モデルの使用は認識性能の向上に効果があり，英語音素モデルとの併用はさらに認識性能を向上させることができ，両音響モデルをタスク適応及び話者適応した場合にも同様の効果があることを示した[6]．本稿では，日本人話者の英語発声の習熟度別の日本語及び英語音素モデルの併用効果について報告する．また，日本人話者の日本語音声を用いて話者適応した場合の効果について検討する．

2．英語音素及び日本語音素モデルの作成

2.1　音素表記

本検討では、英語の単語辞書の音素記述として，英語音素表記(例 probably:*p r aa b ax b l iy*) と日本語音素表記(*p u r o b a b u r i i*)の二種を行った．日本語表記は日本人の典型的な英語発声にあわせて作成した．発声の多様さに対応させるために複数読みを持つ単語は英語音素表記よりも多い．

2.2　評価及び適応用データ

日本人英語音声の評価，適応を行う音声は LDC North American News Corpus より英語音素バランスを考慮して選択した 1000 文を発声したものである(Data-1J)．収録は騒音のない室内で行い，学習・評価に使用した帯域は 100Hz～5.2kHz，男女はそれぞれ 100 名で 1 名当り 100 文を発声している．そのうち約 80 文を評価用(1J-T)とし約 20 文を話者適応用(1J-A)に分割した(これにより各話者の評価音声の perplexity をほぼ同じにした)．タスク適応では，同じ発声内容の音声は学習に含まれない．

比較のために，Data-1J と同じ発声内容を日本在住の米国もしくはカナダ国籍の母国語話者(男女各 50 名)が発声したデータベースを準備した(Data-2E)．また，日本語発声による適応の効果を検証するために，Data-1J と同じ話者が日本語文各 50 文を発声したデータベースを準備した(Data-3J)．

2.3 英語及び日本語音素モデルの作成

音響モデルは 3 状態 8 混合のコンテキスト依存不特定話者モデル（性別非依存）を用いており，英語音素モデルは 48 音素で構成し，日本語音素モデルは 31 音素とした．日本語の音素と英語の音素は別のモデルとし(例えば英語の *p* と日本語の *p* は別のモデル)，無音のみ共通のモデルとした．学習及び適応データを表 1 に示す．

音響モデルは以下の手順で作成した．まず，基本モデルとして英語音素モデルは母国語話者の発声した英語音声，日本語音素モデルは日本人の発声した日本語音声で作成した．次に、評価音声と同じタスクの音声で適応を行い(英語音素モデルは母国語話者英語音声(Data-2E)で，日本語音素モデルは日本人話者英語音声(Data-1J)で適応)，タスク適応モデルを作成した．最後に評価話者と同じ発話者の音声で話者適応を行い(同じ音声データ(1J-A/2E-A)を英語音素モデルは英語音素表記で，日本語音素モデルは日本語音素表記で適応)話者適応モデルを作成した．ここで適応方式は MAP 推定を用いており，タスク適応では，各分布の平均値，分散，重みを適応しており，話者適応では平均値のみを適応している．

表 1　各モデル毎の学習適応データ

学習	適応	内容	量
英語音素モデル	基本	LDC Resource Management & Wall Street Journal	約 49k 発声
	タスク適応	母国語話者英語音素バランス文	10k 文 (男女各 50 名)
	話者	英語音素バランス文	約 20 文
日本語音素モデル	基本	日本語バランス単語，文，外来語	127k 発声．約 400 名
	タスク適応	日本人話者英語音素バランス文 (Data-1J)	19k 文 (男女各 50 名)
	話者	英語音素バランス文	約 20 文

3．比較する認識方式

[方式 1:M1] 英語音素モデルと英語音素表記辞書のみ用いる認識方式

[方式 2:M2] 日本語音素モデルと日本語音素表記辞書のみ用いる認識方式

[方式 3:M3] 方式 1 の認識結果のスコアと方式 2 の認識スコアの高い方を認識結果とする．処理量は方式 1 と方式 2 の和となる．

[方式 4:M4] 辞書に英語音素表記及び日本語音素表記を併記した方式．これにより単語単位にスコアの高い表記を認識結果とできる．

[方式 5:M5] 英語音素表記及び日本語音素表記をそれぞれ別の単語として単語辞書に登録する認識方式．認識結果の連続する単語の読みが同一言語でない場合にスコアにペナルティ（バックオフ値）を与えることになり言語間での頻繁な単語単位の読みの遷移を防ぐ効果がある．

4．話者適応結果とその効果

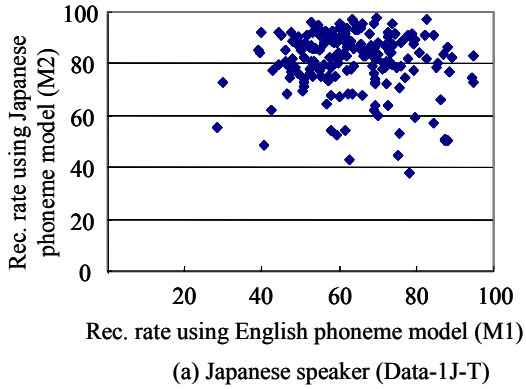
4.1 モデル別認識性能

†日本電信電話株式会社　NTT サイバースペース研究所
NTT Cyber Space Laboratories, NTT Corporation
1-1 Hikarinooka Yokosuka-Shi Kanagawa 239-0847 Japan

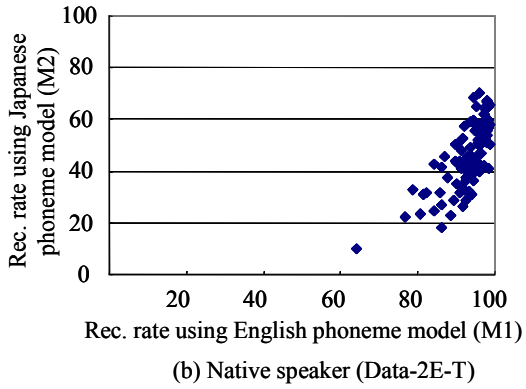
日本人の英語音声(1J)及び、母国語話者の発声(2E)に対して基本モデル、タスク適応モデル，話者適応モデルを用いた認識結果(NTT 音声認識エンジン VoiceRex による)を表 2 に示す．表 2 の話者適応において，(E)は英語文発声のみで適応した結果であり，(EJ)は英語と日本語文の発声により適応した結果である．この結果，日本語文発声を加えて話者適応を行った音響モデルを用いる場合でも，他の学習段階の音響モデルを使用した場合と同様に，①日本語音素モデルの使用が有効であり(M1 < M2)，②英語音素モデルの併用はより効果がある(M2 < M3,M4,M5) ことがわかる．また，頻繁な単語単位の読みの遷移を防ぐことが望ましい(M4 < M5)．図 1 に話者適応後(英語発声で適応した場合)の，日本語音響モデルのみを用いた場合の認識性能と、英語音響モデルのみの認識性能を示す(各点が各話者に対応)．話者適応を行った後においても，母国語話者と日本人の話者の傾向は大きく異なることが分かる．

表 2.　日本語音素モデル使用の効果(%correct)

Model	Data	M1	M2	M3	M4	M5
Baseline	1J	29.1	49.2	52.7	52.8	53.9
	2E	76.2	12.1	73.9	61.8	70.5
Task adaptation	1J	46.8	74.2	78.1	76.9	80.0
	2E	91.4	28.2	91.3	85.6	92.8
Speaker(E) adaptation	1J	63.5	81.7	84.1	85.0	86.4
	2E	93.0	46.1	92.5	87.3	94.2
Speaker(EJ)	1J	63.5	82.7	84.7	83.4	87.1



(a) Japanese speaker (Data-1J-T)



(b) Native speaker (Data-2E-T)

図 1　英語及び日本語音素モデルによる各話者の英語文による話者適応後の認識精度の関係

4.2 日本人話者の英語習熟度別の認識性能

発声者の英語習熟度に応じた認識精度の検証を行うため，日本人の話者を二名の評価者(米国人，及び英語に堪能な日本人)により 5 段階の主観評価を行い，評価値の平均(r)で 4 グループに分類した(r が高いほどより母国語話者に近い発声)．タスク適応，話者適応モデルを用いた認識結果を表 3 に示す．この結果(M1,M2)より，習熟度評価値の高い話者は英語音素モデルを使用することで，評価値の低い話者は日本語モデルを使用することでより高い性能を達成できることがわかる．タスク適応及び，英語発声文で話者適応を行った場合(E)には，すべてグループで精度向上の効果を確認できる．一方英語及び日本語文で適応を行った場合(EJ)には，習熟度の高いグループで，日本語及び英語音素モデルを用いる場合(M3,M5)に英語文のみで適応する場合より精度向上が低い．即ち，母国語話者に近い発声をする日本人話者には英語文のみで話者適応を行う方が良いことがわかる．

表 3　英語発声習熟度別の実験結果

Subject	Adaptation \	M1	M2	M3	M5
$r\leq 2$ (13%)	Task	34.0	82.9	83.1	85.8
	Task+Speak(E)	55.3	87.7	87.7	90.0
	Task+Speak(EJ)	55.3	88.9	88.9	91.7
$2<r\leq 3$ (45%)	Task	39.3	79.9	80.2	82.8
	Task+Speak(E)	58.2	85.6	85.8	88.4
	Task+Speak(EJ)	58.2	86.9	87.0	90.5
$3<r\leq 4$ (30%)	Task	54.6	69.4	73.8	75.4
	Task+Speak(E)	68.9	78.5	81.2	83.1
	Task+Speak(EJ)	68.9	79.1	81.4	84.3
$4<r$ (12%)	Task	71.1	54.4	75.3	73.9
	Task+Speak(E)	79.5	67.7	80.6	83.3
	Task+Speak(EJ)	79.5	68.6	79.3	75.1

5．まとめ

日本人の英語文発声の認識に対して，日本語表記及び日本語音素モデルを用いる効果及び音素モデルの話者適応を行う効果について発声者の英語習熟度に応じて検証を行った．今後はペナルティの適切な設定方法について検討する．

参考文献

[1] Witt, S. and Young, S., "Off-line acoustic modeling of non-native accents", *Proc. EuroSpeech'99*, pp.1367-1370, 1999.

[2] He, X. et al., "Fast model adaptation and complexity selection for non-native English speakers", *Proc. ICASSP-02*, 2002.

[3] Fisher, V., et al., "Likelihood combination and recognition output voting for the decoding of non-native speech with multilingual HMMs", *Proc. ICSLP-02*, pp.489-492, 2002.

[4] 阿部 他, "日本人英語の特性に基づく音声認識を用いた英会話学習支援システム," 音講論, 3-1-9,2001-10

[5] 倉田 他., "発音習熟度に着眼した適応処理に基づく非母国語音声認識の高精度化," 信学技報 S2002-38, 2002

[6] Matsunaga, S., et al., "Non-native English speech recognition using bilingual English lexicon and acoustic models", *Proc. ICASSP-03*, pp.I-340-343, 2003