社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

日本人英語発声に観測される発音上の癖を考慮した音声認識

大崎　功一†　　峯松　信明†　　広瀬　啓吉††

† 東京大学大学院 情報理工学系研究科
†† 東京大学大学院 新領域創成科学研究科
〒 113-0033 東京都文京区本郷 7-3-1
E-mail: †{koichi,mine,hirose}@gavo.t.u-tokyo.ac.jp

あらまし　昨今の国際化に伴い、国際会議など国際標準語である英語を用いて意志疎通を図らなければならない場面が急増してきている。このような場面において認識技術の応用を考えた場合、非母国語の認識技術の向上が急務である。非母国語音声認識の性能向上は、音響モデリング、発音辞書、言語モデリング、デコーディングの各処理系にて検討する必要があるが、本研究では音響モデルに焦点を絞って、その性能向上を目指す。我々は以前、習熟度を考慮して各話者の発声に内在する音声学的構造を推定し、それに基づいてモデル適応する方法について提案した。しかし、習熟度の低い話者に対して適用できない、モデル構築の際に導入される音声学的構造に着目していない、という問題点があった。そこで本研究では、日本人英語モデルと日本人日本語モデルをマルチパス化する手法と、日本人英語に現れるスペルに依存した発音の様子をモデル構築に組み込む手法について検討した。
キーワード　非母国語音声認識、日本人英語、発音習熟度、マルチパスモデル、スペル

Speech recognition of Japanese English using Japanese specific pronunciation habits

Koichi OSAKI†, Nobuaki MINEMATSU†, and Keikichi HIROSE††

† Graduate School of Information Science and Technology, University of Tokyo
†† Graduate School of Frontier Sciences, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan
E-mail: †{koichi,mine,hirose}@gavo.t.u-tokyo.ac.jp

Abstract　Today's globalization often requires Japanese people to talk in English, which is often seen in international meetings and so on. To apply speech recognition techniques into these situations successfully, improvement of non-native speech recognition has to be realised. The improvement can be achieved by enhancing acoustic models, pronunciation lexicon, language modeles, and/or decoding stragegies independently. In this paper, a research focus is put on the acoustic models. In our previous study, we proposed a technique to enhance MLLR-based adaptation with respect to pronunciation proficiency of the speaker, where proficiency-based phoneme regression trees were generated and used in the adaptation. But this method had several drawbacks. It cannot be applied to speakers with lower proficiency or give any solutions to a problem that the commonly-used top-down state-tying may introduce inadequate phonetic structure into the model set. In this paper, we firstly analyze Japanese English and propose a technique of combining Japanese English acoustic models with Japanese ones. Next, we investigate a new state-tying technique which considers spellings of vowels as well as left and right phoneme context.
Key words　non-native speech recognition, Japanese English, pronunciation proficiency, multi-path models, spell

1. 研究の背景と目的

近年の計算機性能の向上及び音声言語情報処理技術の進展により、読み上げ音声に対する大語彙連続音声認識性能は理想的な環境下であれば単語誤り率を 5 %以下にまで低減できるようになった [1]。これらの技術的発展を受け、音声認識のターゲットは背景雑音が混入するような実環境下での認識 [2]、あるいは発話形態として独話音声や対話音声など、より自然な発話の認識へと移りつつある。特に独話音声に対しては大規模な音声データベースが構築されつつあり、独話音声の認識及び要約に対して数多くの研究が行なわれるようになった [3]。聴覚障害者が参加するような会議では、発表者の発話内容を速記者が書き取り、それを文字情報としてスクリーン表示するなどのサービスが行なわれているが、バリアフリー環境の実現を考えた場合、独話音声認識の主要な応用場面である。

一方昨今の急速な国際化により、日本人が英語（すなわち非

母国語）を話さなければならない機会はますます増えてきている。例えば国際会議などを考えた場合、発表者は英語での発表を義務づけられているため、発表者の過半数が非母国語による発表を行なっている。このような場面において上記した独話音声認識の応用を考えた場合、非母国語の音声認識技術は必須である。また、非母国語としての英語を用いて意志疎通を図らなければならない場面は国際会議のような発表の場のみならず、外国からの出席者がいる会議・打ち合わせや対話式の語学学習システム [4]、更には、航空機の管制システム運用 [5] など種々の場面を容易に想定することができる。

非母国語の音声認識は、発音の曖昧さや混同、外来語の影響から来る発音誤り等が引き起こす音響パラメータ分布の歪みや、不適切な語彙の利用等が引き起こす言語パラメータの歪み等、種々の特有の問題が存在する。これらの問題を解決する場合、現在の音声認識の実現形態においては、音響モデリング [6]、発音辞書の実装方式 [7]、言語モデリング、デコーダの実装方式 [8] のいずれか（あるいはその組合せ）において問題解決を試みることになる。本研究ではこれらのうち、音響モデルに焦点を当て、日本人による英語音声の認識精度の向上を図る。

文部科学省科学研究費特定領域研究において、男女 100 人ずつの日本人英語データベースが構築されており [9]、これを用いることによって、日本人英語音響モデルを構築することが可能である。しかしながら、日本人英語の特徴の一つとして発音習熟度の差異に起因する音響パラメータ分布の広がりが挙げられ、構築される日本人英語音響モデルは（データベース中の）平均的な習熟度における音響モデルと考えることができ、習熟度の高い話者・低い話者には必ずしも適合しない。

我々は以前、習熟度を考慮して各話者の発声に内在する音声学的構造を推定しそれに基づいてモデル適応する方法について提案した [10]。しかし、習熟度の低い話者に対して適用できない、モデル構築の際に導入される音声学的構造に着目していない、という問題点があった。そこで本研究では、日本人英語 (JE) モデルと日本人日本語 (JJ) モデルをマルチパス化する手法と、日本人英語に現れるスペルに依存した発音の様子をモデル構築（状態共有時）に組み込む手法について検討した。

2. モデル構築における母国語入力依存性

従来の認識技術構築は、話者が母国語を話す場面のみを想定して構築されてきた。このような技術を非母国語音声認識に用いる場合、各技術構築において、結果的に「認識対象が母国語音声である」という仮定が置かれてしまっているか否かを見極めることは重要である。ここでは、標準的な音響モデル構築手法である「状態共有の triphone モデル」を例にとり考察する。

音素環境利用の是非　日本人英語には、英語単音以外にも日本語独自の単音が混入する。これに対して（母語話者）英語音響モデルのみに基づいて技術構築をする場合、用意された音素セットでは表現できない音が存在することになる。また、日本人英語音響モデルを構築したとしても、日本人英語には発音習熟度に起因する音響パラメータの広がりがあるため、習熟度の低い話者ではさらに日本語独自の単音が混入する可能性がある。このように、非母国語音声認識の向上を考える場合は発音習熟度に起因する影響を深く考慮する必要がある。

状態共有による構造の導入　triphone モデルは、パラメータ数の増加を抑えるために、状態や分布の共有が行なわれる。中でも、音素環境に関する質問セットによる決定木を用いた、トップダウン状態共有が広く行なわれている。この場合、中心音素、及び、HMM 状態位置毎に、前後の音素環境に着眼した状態レベルの決定木が生成される。「状態を共有する」ということは、音素モデルセットにある種の構造を導入することを意味し、この構造は、話者適応処理を事後的に施す場合でも “不変” である。この先天的音響モデルセット構造が「母国語話者らしさ」を反映している場合、非母国語音声認識の性能向上を抑制することが推測される。例えば、日本人は前後音素環境以外に、スペルに依存した発声をしていることが容易に想像されるが、そのような枠組みは従来の英語音声認識では議論されていない。

適応処理における構造の導入　少量の適応データで高い効果を出す適応手法として MLLR が広く使われている。この手法は、不特定話者音響モデルにおいて、ボトムアップ的に混合レベルの回帰木（分類木）を求め、モデルセット内の全混合をクラスタリングする。この場合、状態共有時とは異なり、同一ノードに異なる音素の状態・分布・混合が同居する場合も生じる。このクラスタリングは不特定話者モデルを用いて行なわれるため、それが母語話者モデルであれば、母語話者発声における音声学的構造が直接的に反映される形となり、適応処理の効果を低減させることが容易に想像される [6]。

状態・混合クラスタリングは、如何に高効率な学習・適応を実現するか、を目的として提案されており、そのために対象とするデータに内在する音声学的構造を推定し、利用する、という方法論に基づいている。当然この構造は、元データを発声した話者群に特有の発声構造を反映しているため、母語話者音響モデルをベースとした非母国語音声認識は、この制約（障害）下での議論となり、自ずとその限界が予測される。また、日本人英語モデル（即ち日本人英語の平均モデル）をベースとした場合でも、入力話者の発音習熟度の多様性に十分に追従できない可能性がある。我々は以前、MLLR 話者適応時の混合クラスタリングによる構造導入に対して実験的に検討した。それに対して本稿では、音素環境の問題に関しては日本人英語モデルと日本人日本語モデルをマルチパス化する手法を検討し、状態共有による構造の導入としては、triphone モデル構築時（状態共有時）に母音のスペルを考慮する手法を検討した。

3. 日本人英語分析

日本人英語は、発音が曖昧であったり、音を日本語と混同したり、外来語の影響による発音誤りなどの日本人英語特有の発音誤りが存在するため、母国語英語音響モデルを用いるとミスマッチが起きてしまい、正しく認識することが困難である。しかし、前述したように、現在では、日本人英語データベースを用いて日本人英語音響モデルを作成することが可能である。しかしながら、日本人英語には更に習熟度という軸が存在し、習熟度によって発音誤りの程度が異なる。すなわち、日本人英語

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

デコーダ	Julian rev3.3p2
分析音声	英語学習者音声データベース
	男性話者音声 (87 話者 720 文)
分析窓	16bit/16kHz サンプリング
高域強調	$1 - 0.97z^{-1}$
特徴量	12MFCC + 12Δ MFCC + Δ Power
日本人英語 HMM	日本人英語データベースから
	構築した monophone(混合数 16)
	5 状態 3 分布 (但し sp のみ 3 状態 1 分布)
日本語 HMM	JNAS の monophone(混合数 16)
	5 状態 3 分布 (但し sp のみ 3 状態 1 分布)

表 2　日本語及び英語音素セット
Table 2　Phoneme sets of Japanese and English

英語の母音	ae,ah,eh,ih,oy,er,uh,aw,
	ay,aa,ao,ey,iy,ow,uw,ax
英語の子音	ch,dh,nx,jh,sh,th,zh,b,d,
	f,g,hh,k,l,m,n,p,r,s,t,v,w,y,z
日本語の母音	a,i,u,e,o
日本語の子音	N,b,ch,d,f,g,h,j,k,m
	n,p,r,s,sh,t,ts,w,y,z

音響モデルはあくまで「日本人の平均英語音響モデル」と考えることができ、例えば習熟度の低い話者に関してはミスマッチが多発していることが考えられる。そこで、日本人英語に予測される誤りを許可するようなネットワーク文法を使用し、日本人英語 monophone と日本人日本語 monophone を用いて音素単位の認識を行なうことにより、日本人英語分析を行なった。分析条件を表 1 に、分析に用いた日本人英語と日本人日本語の音素セットを表 2 に示す。

構築したネットワーク文法は以下の通りである。

a） 母音挿入誤り

/C/(C:日本人英語子音。n,w,y は除く) が単語尾又は後続音素が子音の場合、当該子音の直後にあらゆる日本人日本語母音が挿入されてもよい。

b） 置 換 誤 り

各音素は日本人にとって似ているか、調音位置の似ている日本人日本語音素に置換されてもよい。また、母音はスペルからくる発音誤りも考慮する。

c） 脱 落 誤 り

/V r/, /V l/, /V h/(V:母音) が単語尾又は後続音素が子音の場合当該/r/, /l/, /h/が脱落してもよい。

日本人英語分析を行なった結果を図 1 に示す。母音挿入誤りと脱落誤りに関しては共に誤り率が低かったため、ここでは置換誤り分析結果のみを載せる。日本人英語音響モデルを用いているにも拘わらず、音素によっては日本人日本語音素への置換率が 50%を超えるものも存在する。このことは、日本人英語の認識に関して日本人英語モデルをそのまま用いるのでは不十分であることを示唆している。また、習熟度の低い話者のみを集めた場合には、置換率は更に増加することが考えられる。

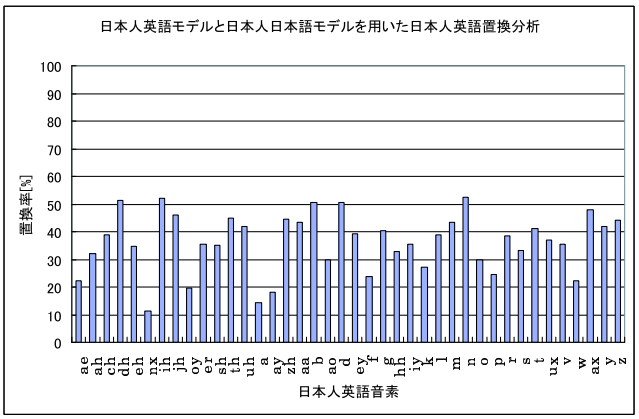

図 1　置換分析結果
Fig. 1　Results of analysis of phoneme replacement

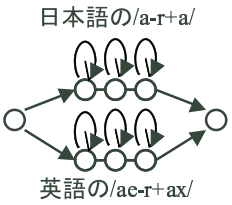

図 2　マルチパス置換モデル
Fig. 2　An example of multi-path models

4.　JE+JJ マルチパス置換モデル

4.1　マルチパス置換モデルの構築

図 2 のように日本人英語モデルに日本人日本語モデルを組み込むことにより、認識率の向上が期待される。マルチパス置換モデル構築の際に最も問題となるのは、日本語音素と英語音素の対応とマルチパス時の分岐確率である。これらの問題は日本人英語モデルと日本人日本語モデルを用いた日本人英語分析から、各音素に対して最も置換の多かったものを取得し、それを対応リスト (表 3) とすることで monophone の対応をとり、これを利用するなどで解決を図った。具体的なマルチパス置換モデル構築の流れは以下の通りである。ここで、日本人日本語モデルへの分岐確率を置換率と呼ぶこととする。

- 各日本人英語 triphone に関して前後の音素コンテキストも含めてリストにある日本語音素に変換する。
- 変換された日本語 triphone が存在すれば当該日本語 triphone へのマルチパス置換モデルを構築する。なければ前 biphone、後 biphone、monophone の順に変更し、日本語内に存在する音素へのマルチパス置換モデルを構築する。
- 全ての日本人英語音素に対して置換率が p のモデルを構築する（p=0.0,0.1,0.2,$\cdots$,1.0、合計 11 個)。$p = 0.0$ とは日本人英語モデルのことである。
- 得られた各モデルに対して適応文の適応文の forced alignment をとり、各中心音素に対してフレーム正規化尤度が最も高くなる置換率を取得する。
- 日本人英語モデルの正規化尤度と比較して尤度の上がり幅が閾値よりも高い音素に関してのみマルチパス化する。

表 3　日本人英語と日本人日本語の対応リスト (一部)
Table 3　Correspondence between JE phonemes and JJ ones

JE	JJ	JE	JJ	JE	JJ	JE	JJ
ae	a	ah	a	ch	ch	dh	z
eh	e	nx	n	ih	e	jh	j
oy	o i	er	a	sh	sh	th	s
uh	u	aw	a u	ay	a	zh	sh

表 4　認識実験の条件
Table 4　Conditions of recognition experiments

デコーダ	Julius rev3.3p2-multipath
音響モデル	状態数 2000 の triphone
言語モデル	前向き bigram、後向き trigram
語彙数	20K
適応文	平均 100 文/話者
評価文	20 文/話者（未知語率 8%）
PP.	前向き bigram に対して 500 以下

表 5　置換モデルによる認識率 [%]
Table 5　Recognition rates using the multi-path models

話者	GOP	JE	JE+JJ
話者 A	-1086.25	62.26	62.26
話者 B	-904.88	67.52	68.15
話者 C	-800.35	74.50	74.50
話者 D	-749.12	72.03	77.12
話者 E	-689.38	64.78	66.04
話者 F	-600.19	65.13	67.76
話者 G	-550.26	66.14	66.14
話者 H	-505.33	66.45	64.47
話者 I	-447.81	63.95	61.22
話者 J	-195.04	54.78	52.23

閾値は実験的に決定し、ここでは 0.3 とした。

4.2　認識評価実験

提案手法の評価実験を 10 名の評価用日本人話者を用いて行なった。認識実験条件を表 4 に示す。比較実験としては、日本人英語モデルに関して検討した。実験により得られた単語正解率を自動推定された GOP(Goodness Of Pronunciation、話者習熟度 [11]:値が大きい方が習熟度が高い) と共に表 5 に示す。

日本人英語モデルと比較して、習熟度の高い話者 H,I,J に関しては認識率の減少が見られる。これは、習熟度の高い話者は日本人英語モデルから日本人日本語モデルへの置換が少ないと考えられるため、このマルチパス置換モデルは有効に働かなかったものと考えられる。他の話者については、認識率の向上が見られた話者と、認識率の変わらない話者、というのが存在する。これは、恐らく英語 triphone と日本語 triphone の対応が一対一で取れないために、よほど置換の多い話者でなければマルチパス置換モデルが有効に働かないためだと推察される。

今後の課題としては、より良い英語 triphone と日本語 triphone の対応づけと、コンテキストに依存した置換率の制御があり、それらを組み込むことによって、更なる性能向上が期待できる。

表 6　スペル母音例（出現頻度順）
Table 6　Examples of extended vowels with their spells

元母音	スペル母音（“_”以降がスペルである）
ax	ax_a, ax_e, ax_o, ax_u, ax_i, ax_ou, ax_?
eh	eh_e, eh_ea, eh_a, eh_ai, eh_?
ih	ih_i, ih_e, ih_io, ih_a, ih_y, ih_o, ih_ui, ih_ee, ih_ia, ih_?
ay	ay_i, ay_y, ay_ie, ay_ui, y_uy, y_y, y_?

5.　スペル情報を考慮した状態共有モデル

第 2. 節において、一般的な前後音素に着眼した状態クラスタリングでは、日本人英語特有の発音形態を十分に反映できない可能性があることを示した。その代表例がスペルに基づく発音の偏りである。例えば PROLEX 辞書では、above, useful, common は同一の音素 /ax/（弱母音、schwa）が割り当てられている。耳から英語を獲得した母語話者はこれらの音を区別することは無いが、目から英語を習得した日本人には、これらの音を区別せずに発音することの方が困難である。

5.1　音素とスペルの対応付け

triphone 学習における状態共有において、前後音素のみならず、スペルを考慮する場合、音素とスペルの対応を正確にとる必要がある。しかし、英語という言語は、表記と発音とが一致しない言語であり、その対応を求めるのは困難である。そこで、母音のみに着眼し、対応を以下の方法で求めた。なお母音に着眼した理由は、両言語を比較した場合、母音体系の方が子音体系よりも言語間差異が大きいからである。

- 着目する単語の音素記号列から母音数 (N_v) を取得する
- その単語のスペルから、文字 `[aiueo]` のいずれかのみで構成される部位（母音部位）とそれ以外（子音部位）に分割する。文字 `e` のみで構成される部位数 (n_e) を取得する
- 子音部位に文字 `y` が出現し、その前後に `y` 以外の文字が存在する場合、その `y` も母音部位として新たに登録する
- N_v= 母音部位数の場合、母音部位と母音の対応をとる
- N_v= 母音部位数 $-n_e$ の場合、`e` を除いて対応をとる
- 母音が二重母音の場合、対応する母音部位に後続する文字が `y` 或いは `w` の場合は、その文字も母音部位に含める
- 対応がとれない母音は「スペル不明」とする

本手順を DB 中の全文に対して行なったところ、約 90%の母音がスペルと対応し、内 100 文に対して対応精度を求めたところ約 95%の精度を得た。

5.2　スペル triphone モデルの学習

スペルとの対応がとれた母音をスペル母音として定義し、triphone を学習する。この際、出現頻度の低いスペル母音は「スペル不明」母音として学習した方が有利である。そこで、スペル対応がとれた母音を集計し、母音別に累積出現回数が 98%を越えるまで対象母音のスペル母音を定義し、それ以降は「スペル不明」として定義した。その結果母音数が 16 から 95 へ増加した。表 6 にスペル母音例を示す。

triphone モデル学習において状態共有を実現する場合、状態を分割する際に試行する質問セットを用意する。通常は前後音

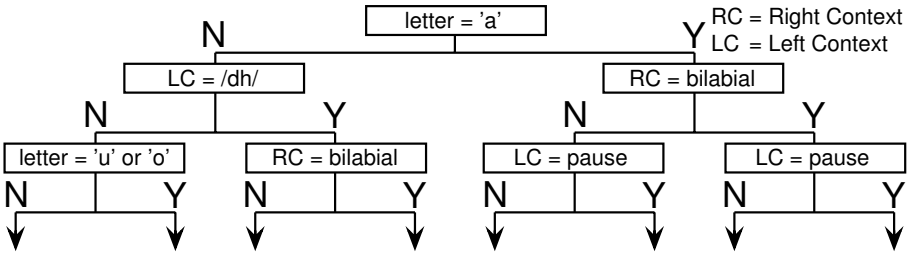

図 3　/ax/中心状態に関する状態分割決定木

Fig. 3　A part of a decision tree trained for the central state of /ax/

素環境に対する質問が用意されるが、ここでは、スペルに着眼して状態を 2 分割するための質問を追加する必要がある。ある母音に対してスペル母音及びスペル不明母音が合計 n 種類あった場合、スペルに基づく 2 分割は合計 $(2^n - 2)/2$ 通りある。これだけの規則を各母音に対して追加することになる。その結果、通常の状態共有では 118 個であった質問が、スペル母音導入後、1414 個まで増加した。

決定木構築後、構成された木においてスペルに関連する質問がどのように利用されているのかについて検討した。16 種類の米語母音中、6 割以上の母音中心状態においてスペル質問が採用されていた。それらの全母音において、木の深さ 3 以下の個所においてスペル質問が採用されていた。これらは、スペルに着眼した状態分類の有効性を意味するものである。図 3 に /ax/ (schwa 音) の中心状態に対する決定木を示すが、root node における第一分割においてスペルが参照されている様子が分かる。

5.3　学習データの少なさを考慮したスペル母音選択

5.2 節においてスペル triphone モデルを構築したが、母音数が 16 から 95 へと大幅に増加している。それに対して日本人英語データベースに存在する日本人英語音声データは約 10,000 文程度と必ずしも多くない。そのため、第 5.2 節のような枠組みでスペル triphone を構築した場合、各スペル triphone に対する学習データ量が大幅に減少してしまう可能性がある。特に top-down クラスタリングを用いた決定木構築では、(学習データ中の) 異なり論理 triphone を、一端、異なり物理 triphone として構築する必要があるため、この時点でデータ量の低減は性能劣化に結びつく可能性が高い。そのため、有効でないスペル母音は出来る限り除いた方が望ましいと考えられる。

そこで、次の手法により有効なスペル母音のみ定義し、そうでないスペル母音はスペル不明母音としてまとめて定義する。

（1） スペル triphone モデル構築の際に得られる状態共有決定木を参照し、スペルに関する質問のみを抽出する。

（2） 質問を各スペル母音毎に分け、登場回数を数える。

（3） 決定木に登場した各スペル母音を登場回数の多い順に並べ、登場回数の多いスペル母音からその累積登場回数が閾値 θ %以上になるまで新たなスペル母音として定義し、それ以外はスペル不明母音とする。

（4） 5.2 節の手法を適用し、スペル triphone を再構築する。

本研究では、θ として 100、80、50 とし、そのそれぞれに対してスペル triphone を構築した。ここで、5.2 節で作成したスペル triphone モデルと θ を 100 として構築したスペル triphone は異なるモデルである。何故ならば、前者は第 5.1 節において決定したスペル母音を全て用いているのに対し、後者はその中で、状態共有決定木に登場したスペル母音を用いているのであって、登場していないスペル母音は除いているからである。

表 7　スペル triphone 認識実験の条件

Table 7　Experimental conditions with the extended vowels

デコーダ	Julius rev3.3p2
音響モデル	状態数 1000 の triphone
言語モデル	前向き bigram、後向き trigram
語彙数	20K
評価文	open 話者 32 名 980 文（Task-A）
	（未知語率 8%, PP.200 以下）
	open 話者 10 名, 20 文/話者（Task-B）
	（未知語率 8%, PP.500 以下）

5.4　認識評価実験

スペル triphone モデルの評価実験を行なった。認識実験条件を表 7 に示す。表 7 にも示した通り、2 種類の評価データを用いて評価を行なった。1 つは open 話者 32 名 980 文のデータを一括して認識したもの。もう 1 つはマルチパス置換モデルと同様の、GOP 毎に分けられた 10 名それぞれ各 20 文ずつの評価データを用いて話者毎に認識を行なったものである。

比較実験としては表 8 に示すように、日本人英語 triphone モデル (JE)、学習の際にはスペル母音を考慮し状態共有の際にはスペルを考慮しないモデル (no-spell)、通常のスペル triphone モデル (S-full)、θ を 100、80、50 としてスペル母音を定義したスペル triphone モデル (S-100、S-80、S-50) に対して認識実験を行なった。ここで、no-spell とは、学習の際には S-full と同様にスペル母音を定義して学習を行なう。しかし、状態共有の質問セットにはスペルに関する質問を入れない、としたモデルである。つまり、最終的にはスペルを考慮しない日本人英語 triphone が出来上がることになる。このモデルは、スペル化する際の各スペル母音に対する学習データの少なさのみを反映させた日本人英語 triphone ということができる。すなわち、no-spell と S-full を比較することで、「学習データの少なさ」という要因を除くことが出来ると考えられる。

認識実験の結果を表 9、10 に示す。まず表 9 について考察する。表 9 を全体的に見て、スペル化したモデルの方が良い認識率が得られており、このことからも、日本人英語はスペルに依存した発声をしており、それを音声認識の枠組みの中に組み込

表 10　スペル triphone 認識実験結果 (10 名)[%]
Table 10　Recognition rates for task-B

話者	GOP	JE	no-spell	S-full	S-100	S-80	S-50	最高認識率
話者 A	-1086.25	67.3	66.7	65.4	69.2	64.2	65.4	S-100
話者 B	-904.88	68.8	68.8	69.4	72.0	72.0	69.4	S-100 S-80
話者 C	-800.35	74.5	67.1	69.8	69.1	71.8	73.2	JE
話者 D	-749.12	76.3	73.7	78.0	78.0	74.6	74.6	S-full S-100
話者 E	-689.38	67.3	64.8	66.0	67.3	68.6	68.6	S-80 S-50
話者 F	-600.19	67.1	67.1	62.5	61.2	59.9	66.5	JE no-spell
話者 G	-550.26	66.9	61.4	63.8	63.8	67.7	66.9	S-80
話者 H	-505.33	68.4	62.5	63.2	63.2	62.5	65.8	JE
話者 I	-447.81	68.7	64.6	72.1	77.6	66.7	68.7	S-100
話者 J	-195.04	45.9	38.2	46.5	46.5	48.4	49.0	S-50

表 8　スペル triphone において比較した音響モデル
Table 8　Triphone models used in the experiments

条件	音響モデル
JE	日本人英語 triphone
no-spell	学習の際にはスペル母音を定義するが、 状態共有の枠組みではスペルを考慮しない triphone
S-full	スペル母音としてスペルの対応を 上位 98 %とったスペル triphone
S-100	S-full の状態共有木を参照して質問セットに 現れたスペル母音のみを定義したスペル triphone
S-80	S-100 のうち上位 80%のスペル母音を 定義したスペル triphone
S-50	S-100 のうち上位 50%のスペル母音を 定義したスペル triphone

表 9　スペル triphone 認識実験結果 (32 名 980 文)[%]
Table 9　Recognition rates for task-A

JE	no-spell	S-full	S-100	S-80	S-50
69.5	68.2	70.5	71.5	70.4	70.7

むことが有用であることが分かる。その中でも、S-100 の認識率が最も高く、スペル化母音はある程度絞った方が良い、ということがいえる。また、JE と no-spell では JE の方が認識率が高いため、やはり学習データの少なさがモデルの精度を下げていることが伺える。しかし、no-spell と S-full を比較した場合においては S-full の方が高い認識率を示しており、このことは、学習データ量の増加により、S-full の認識率がさらに向上することを示唆しているといえる。

次に表 10 について考察する。表 10 は、話者を GOP 順に並べ、それぞれの話者に対して認識率をみたものである。最高認識率とは、各話者が最高認識率を出したモデルを示している。表 10 を見ると、10 人中 7 人がスペル化した方が認識率が高い結果が出ており、表 9 と共に、スペル triphone の効果を示している。また、習熟度によってあまり特徴が出ていないことから、この誤りは日本人全般の特徴とも考えられる。

6. ま と め

本研究では、日本人英語の特徴的な誤りに着目し、日本人英語+日本人日本語マルチパス置換モデルの導入及びスペルに依存させた状態共有を導入することで、日本人英語の音声認識性能の向上を検討した。いずれの場合も、全ての話者において性能向上を実現するまでには達していないが、今後、前者においては日本人英語音素と日本人日本語音素との対応、前後音素に依存した分岐確率の導入を、後者においては話者毎に適切な閾値の決定方法と、日本人英語モデル構築の際にアメリカ人英語モデルでスペル triphone を作り、それに対して日本人英語全データを用いて適応をかけるなどの手法を検討している。

文　献

[1] J. B. Allen, “From Lord Rayleigh to Shannon: How do humans decode speech?”, Plenary lecture of ICASSP’2002 (2002)

[2] 伊田他, “雑音 DB を用いたモデル適応 HMM の SN 比別マルチパスモデルによる雑音下音声認識”, 信学技報, SP2001-92, pp.51-56 (2001)

[3] 古井, “話し言葉の音声認識・理解を目指して”, 信学技報, SP2002-48 (2002)

[4] 中川他, “音声認識技術を利用した英会話 CAI システム”, 情報処理学会論文誌, vol.38, pp.1649-1658 (1997)

[5] 鈴木他, “日本人英語の発話様態を考慮した英語音声連続認識の検討”, 音講論, 1-R-17, pp.151-152 (1998)

[6] X. He, *et al.*, “Fast model adaptation and complexity selection for non-native English speakers”, Proc. ICASSP’2002, pp.577-580 (2002)

[7] C. Huang, *et al.*, “Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition”, Proc. ICSLP’2000, pp.838-841 (2000)

[8] N. Binder, *et al.*, “Recognition of non-native speech using dynamic phoneme lattice processing”, Proc. spring meeting of ASJ, 3-P-19, pp.203-204 (2002)

[9] N. Minematsu *et al.*, “English speech database read by Japanese learners for CALL system development,” Proc. LREC2002, pp.896–903 (2002)

[10] N. Minematsu, G. Kurata, and K. Hirose, “Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition,” Proc. ICSLP’2002, pp.529–532 (2002)

[11] 大崎功一、峯松信明、広瀬啓吉 “非母国語音声認識を目的とした語発音モデリングに関する実験的検討” 音講論集, 2-9-22, pp.105–106 (2002.10)

[12] http://www.ldc.upenn.edu/Catalog/LDC97L20.html