日本人による国際標準語（英語）発声を対象とした自動音声認識

峯松 信明*† 倉田 岳人* 大崎 功一* 広瀬 啓吉**

*東京大学大学院情報理工学系研究科
†Royal Institute of Technology, Stockholm, SWEDEN
**東京大学大学院新領域創成科学研究科

1 はじめに

近年高度化する情報・ネットワーク社会において，その社会インフラを利用者の区別なく提供するバリアフリー社会の必要性が叫ばれている。音声情報処理技術を用いた言語（音声）バリアの除去についても検討され，例えば，聴覚障害者を対象とした講演音声の自動字幕化などが考えられている。従来音声認識の対象は母国語を対象とすることが多かったが，例えば国際会議の講演字幕化などを考えた場合，過半数が非母国語の講演となる。現在の音声認識技術では，例えば，日本人英語教師の英語音声に対する（英語母語話者音声を用いて構築された認識装置の）認識性能は抑制されてしまう。即ち，母国語音声認識のみでは，国際的な舞台における講演字幕化の実現は非常に困難である。

大語彙連続音声認識技術は，音響モデル，言語モデル，発音辞書，デコーダの各種モジュールから構成される。本研究では，音響モデルに焦点をあて，日本人によって発声された英語音声認識について検討する。音響モデルによる解決を考える場合，英語母語話者音声によって構築された音響モデルに対して，日本人英語音声を用いて適応をかける方法が考えられる。しかし後述するように，一般的に使われる triphone モデルには，学習話者（母語話者）発声に観測される音声学的「構造」が埋め込まれ，これは適応後も保存される（即ち先天的構造として残る）ため，性能向上に限界が予想される [1]。幸いなことに，約 200 人の日本人によって読み上げられた日本人英語音声データベースが文部科学省科学研究費プロジェクト「メディア教育」において構築されており [2]，本研究ではまず，この DB を用いて音響モデルを構築した。非母国語音声の一つの特徴は，習熟度の違いによる発音歪みの多様性であるが，本研究では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構築，モデル適応する方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を直接利用した場合，用意された音素セットでは表現できない単音が存在することになる。日本語音響モデルを併用することで問題は回避できるが，音素環境依存モデルを対象とした場合，日本語音素に囲まれた英語音素モデル（及びその逆）が必要となる。このようなデータは当然，母国語話者の音声データには存在しないため，結局，日本人の英語音声 DB を用いたモデル学習が必要となる。

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，そ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になる [1]。

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

Automatic Speech Recognition of English Spoken by Japanse
Nobuaki MINEMATSU, Gakuto KURATA, Koichi OHSAKI, and Keikichi HIROSE
Graduate School of Information Science and Technology, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
mine@gavo.t.u-tokyo.ac.jp

表 1: 音響モデル構築条件

AD 変換	16bit/16kHz サンプリング
分析窓	Hamming 窓 (窓長 25msc/シフト長 10msc)
高域強調	$1 - 0.97z^{-1}$
特徴量	12MFCC + 12ΔMFCC + ΔPower
音素体系	ae, ah, ch, dh, eh, nx, wh, ih, jh, oy, er, sh, th, uh aw, ay, zh, aa, b, ao, d, ey, f, g, hh, iy, k, l, m, n ow, p, r, s, t, uw, v, w, ax, y, z, sp, silB, silE
HMM	状態共有 triphone（混合数 16，総状態数約 2,000） 5 状態 3 分布（但し sp のみ 3 状態 1 分布）
学習データ	母国語話者：245 人，25,652 文 日本人話者：68 人，8,282 文
初期モデル	TIMIT（4,346 文）から構築した monophone

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。以下，MLLR 話者適応時の混合クラスタリング，及び，triphone モデル構築時の状態クラスタリングによる構造導入に対して，その解決方法を実験的に検討した。

3 不特定話者日本人英語音響モデルの構築

男声音声試料を用いて，混合数 16，総状態数 2,000 の状態共有 triphone を学習し，これをベースラインモデルとする。学習条件を表 1 に示す。また，同一条件下，WSJ1 中の男性母語話者音声を用いた英語音響モデルも作成した（表 1 参照）。日本人英語データ，母語話者英語データに対する音素ラベルに関しては，PRONLEX[3] を参照して付与した。日本人英語音声中に頻発する単語間のポーズに関しては，モデル構築過程において認識処理を行ない，sp を適宜挿入した。

4 習熟度の自動推定に基づく MLLR 適応

4.1 従来の MLLR 適応の問題点

MLLR 適応は，混合レベルの回帰木（分類木）を構成し，同一ノードに属する混合に対して，適応データの出力確率が最大となるような変換行列を求め，線形変換を施す手法である。図 1 は，日本人英語 monophone より構成した木の様子である。/r/ と /l/ 音の位置など，日本人英語特有の構造が明確に反映されている。このような構造は当然母語話者モデルには存在しない [4]。即ち，母語話者モデルから作成された木に基づく MLLR 適応は，その性能向上に限界があるのは明白である。一方，日本人英語ベースラインモデルに MLLR を適応することを考えた場合，ベースラインモデルによる木は日本人の平均的な英語発音形態を反映した木となる。一方話者の英語習熟度は個人差が大きいため，平均的な木構造を用いた適応処理が，常に入力話者に適したモデル変換になるとは限らない。以下，話者の発音習熟度，及び，話者に適した木構造を推定し，推定された木構造に基づいた適応処理を検討する [5]。

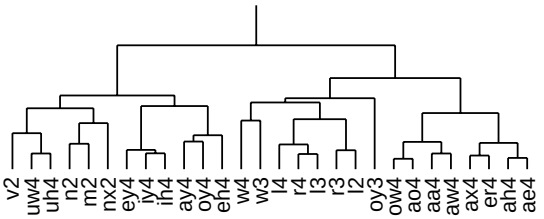

図 1: 日本人英語音響モデルより構成した回帰木（一部）

4.2 内挿モデルを用いた木構造の推定

英語習熟度の推定を内挿モデルを用いて行なう。内挿モデルとは，1 混合 monophone の母国語話者モデルと日本人話者モデル間で，平均，分散，遷移確率の内挿をとって生成したモデルを言う。なお triphone を用いた場合，状態共有の構造がモデルセット間で異なるため，母国語話者モデルと日本人モデル間との対応がとれなくなる。同様に混合モデルの場合も，混合インデックス間の対応が原理的にはとれないために，内挿モデルの生成が困難である。内挿の比率を変化させることで種々のモデルが構成されるが，各々について対象話者の少数音声サンプルの forced alignment を行ない，尤度最大となる重みを決定した。最大尤度重みは，話者によって異なる。この重みの値はその話者の発音が，平均的な日本人英語発音及び，母語話者発音からどの程度離れているのかを示す指標であるため，発音習熟度の指標として捉えることができる。実際に，最大尤度重み位置と英語教師による発音評定点との相関を調べたところ，比較的高い相関値が得られた。と同時に，尤度最大重み時の音響モデルより構成される木構造が，その話者に適した木構造であると解釈できる。

4.3 状態レベル回帰木の triphone への適用

以上の手順により，入力話者に適した，1 混合 monophone から構成した回帰木が構成される。しかしこの回帰木は HMM 状態を単位とした回帰木であり，MLLR が要求する混合レベルの回帰木とは異なる。しかしこの場合「中心音素，HMM 状態位置が等しい混合は全て同一のノードに属する」という制約を課すことで，その適用が可能となる。以上述べてきた提案手法の概要を図 2 に示す。提案手法の利点は，習熟度に着眼した話者適応の実現であるが，図にあるように，日本人英語の状態共有 triphone，1 混合 monophone，母国語話者の 1 混合 monophone など，利用可能な DB から構築される種々の音響モデル（利用可能な種々の情報源）を効率良く活用している点も本手法の特徴と言える。

4.4 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 2 に示す。比較実験としては表 3 に示すように，不特定話者日本人英語モデル，不特定話者母語話者モデルに対して従来の MLLR 適応を行なった場合，行なわなかった場合を検討した。（提案手法を含め）MLLR 適応は，まず全混合に対し

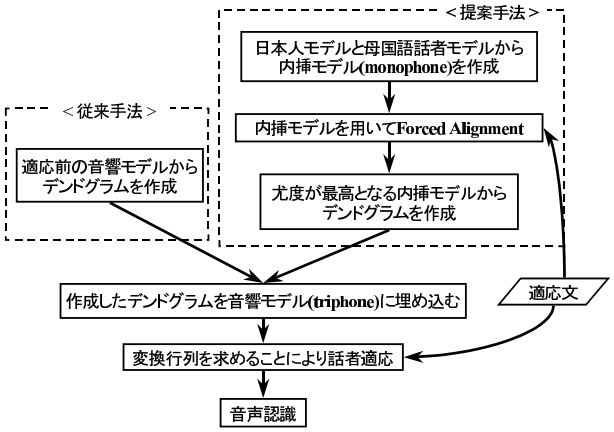

図 2: 提案手法の概要

表 2: 認識実験の条件

デコーダ	Julius rev3.2
言語モデル	前向き bigram，後向き trigram
語彙数	20K
適応文	30 文/話者
評価文	23 文/話者（未知語率 8%）
PP.	前向き bigram に対して平均 260

表 3: 比較した音響モデル

条件	適応前のモデル	適応手法
条件 1	日本人英語モデル	適応なし
条件 2	母国語話者英語モデル	適応なし
条件 3	日本人英語モデル	従来の MLLR
条件 4	母国語話者英語モデル	従来の MLLR
条件 5	日本人英語モデル	提案手法

表 4: 各条件下での単語正解率 [%]

話者	条件 1	条件 2	条件 3	条件 4	条件 5
話者 A	56.2	72.2	66.7	84.9	85.4
話者 B	92.7	44.8	93.2	62.1	93.2
話者 C	91.3	62.1	92.1	72.6	94.1
話者 D	90.9	53.9	94.5	69.4	94.1
話者 E	88.6	47.5	89.5	53.4	91.8
話者 F	89.0	38.8	90.0	48.4	—
話者 G	68.5	40.2	81.3	51.6	—
話者 H	95.4	39.7	95.4	52.5	—
話者 I	89.0	39.3	88.1	52.5	—
話者 K	70.8	42.5	83.6	60.3	—

て共通の変換行列を用いた大局的な適応を行なった後に，回帰木情報に基づく適応を行なった。後者の場合，木の深さは条件 3 における最適値を条件 4，5 に対しても使用しており，条件 5 における最適値を利用することで，提案手法の更なる性能向上が可能である。

実験の結果得られた単語正解率を表 4 に示す。なお，本手法では，内挿モデルによって尤度ピークが観測された話者のみを対象としている。不特定話者日本人英語モデルが，日本人の平均的な英語発音形態を表現しているとすれば，尤度ピークが観測される話者は約半数となることが予想されるが，今回の評価実験でも 10 人中 5 人においてピークが観測された。

本提案手法を適用した 5 人の話者に関して考察する。従来の MLLR を用いた場合の単語誤り削減率は平均 22.0%であったが，提案手法を用いることで平均 30.0%へと改善されている。特に話者 C では，9.2%から 32.2%へと改善され，その効果は非常に大きい。尤度ピークが観測されない残り 5 名について，予備実験的に HMM パラメータの平均，分散に関する外挿を行なって回帰木を構成したが，性能向上は見られなかった。

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，その音素がどのような文字列で表現されているかを考慮する場合，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，母音のみに着眼して音素とスペルの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` 以外の子音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，母音部位と母音との対応をとる。
- 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 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。そこで学習データ中，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」母音として定義した。その結果母音数が 16 から 95 へ増加した。表 5 にスペル母音の例を示す。

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。通常は前後音素環境に対する質問が用意されるが，

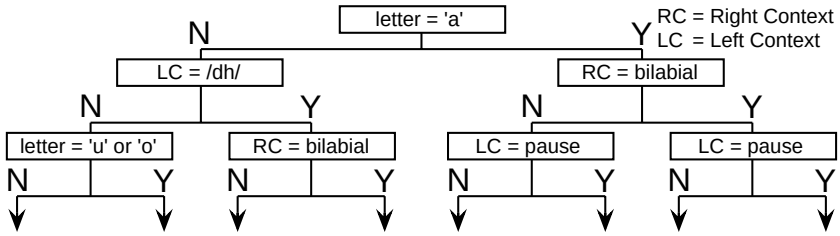

図 3: 構築された /ax/ 中心状態に関する状態分割決定木

表 5: スペル母音例（出現頻度順）

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_?

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。

本手法は，日本人英語特有の発音歪みとして，スペルに基づく発音歪みに着眼し，状態クラスタリング時においてスペルによる状態分類を導入することの是非を検討している。triphone 学習時に構築された決定木の様子を図 3 に示す。図では，音素 /ax/ (弱母音，schwa) の第二状態（即ち中心状態）に対する決定木である。前後音素よりもスペルに対する分割がまず行なわれている。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。ま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直接意味するものである。

5.3 認識評価実験

比較対象として，以下の 2 種類の triphone を用意した。一つは通常のベースライン triphone である。学習において状態共有を行なう場合，学習データに観測される全異なり triphone において，一端，共有構造をもたないモデルを学習する必要がある（その後に，状態共有を行なう）。スペル triphone も同様に，全異なり triphone を一端別個に学習することになるが，その後の状態共有において，スペルが異なる同一母音を共有させ，最終的にスペルを無視した triphone を構成することが可能である。こうして学習される（スペル無し）triphone も比較対象とした。認識実験の結果を表 6 に示す。実験条件などは基本的に表 2 と同一である。

実験の結果，今回構築したスペル triphone の認識性能は，従来の triphone と比較して僅かに低い率を呈した。一方，スペル triphone として一端学習したモデルを状態共有の枠組みの中で再度（スペル無し）triphone として学習して構成されたモデルよりは高い性能を示した。二つのスペル無し triphone において性能差が出るのは，triphone を個別に学習する際の学習データが極端に少なくなるため，学習精度が落ちることが原因であると考えている。スペル triphone も同様の学習過程を経るため，それに起因する精度劣化が付随してしまう。この問題は，スペル母音の種類数を減らすことで対処可能である。木構造に反映されないスペル母音を排除するなどして，意味のあるスペル母音のみに着眼することで性能向上が可能であると考えている。

表 6: スペル triphone を用いた認識実験結果

baseline	スペル triphone	スペル無し triphone
84.7 %	84.2 %	83.6 %

6 まとめ

本研究では，状態共有 triphone モデル構築及び適応の際に導入される音声学的構造に着眼し，入力話者の発音習熟度を考慮することで，その話者に適した構造を推定し，学習及び適応の高精度化を試みた。MLLR 適応に対する実験結果では，入力話者の習熟度を自動推定し，より適した木構造を推定することの効果が得られたが，学習過程における状態共有に母音とスペルの対応を導入する試みに関しては，実験的な性能向上は見られなかった。しかし，学習時に得られる木構造にはスペルに依存した発音歪みが明らかに観測されており，今後，意味のある母音スペル対応を選定するなどして，学習データ量問題の解決を図る予定である。

参考文献

[1] X. He, *et al.*, "Fast model adaptation and complexity selection for non-native English speakers," Proc. ICASSP'2002, pp.577–580 (2002).

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

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

[4] N. Minematsu, G. Kurata, and K. Hirose, "Corpus-based analysis of English spoken by Japanese students in view of the entire phonemic system of English," Proc. ICSLP'2002, pp.1213–1216 (2002)

[5] 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)