Internal Use Only
002

TR-IT-0165

音素クラスタモデルを用いた未登録語検出法の検討
Detection of Unregistered-Words Using Phoneme Cluster Models

坂本 博之 Hiroyuki Sakamoto
松永 昭一 Shoichi Matsunaga

1996.3.28

すべての音素を数種類のクラスタに分類し作成した音響モデル (音素クラスタモデルとよぶ) を用いた未登録語検出法を提案する. ここでは日本語の音節構造を考慮したクラスタモデル, 音響モデルの自動クラスタリングにより決定したクラスタモデル, 全音素を1つのクラスタとしたクラスタモデルの比較検討を行った. 未登録語を含んだ文章の未登録語検出実験において, 日本語の音節構造を考慮した音素クラスタモデルは, 従来の音素毎のモデルを用いる方法と比較して, 処理量を約半分に削減しながらほぼ同等の単語 accuracy を達成することができた. このことから, 提案する方法が処理量を抑えた未登録語検出に有効であることが分かった. またこの時, 未登録語区間のスコアに対してクラスタ N-gram の確率をペナルティとして使用することが, 有効であることが分かった. また, 多段階による認識方式の1段階目として本提案法により未登録語を検出し, 2段階目で検出された未登録語の音素系列を認識する方式についても述べる.

1 はじめに

大語彙連続音声認識システムの多くは，高い認識性能と高速な処理を同時に実現するためにあらかじめシステム内の辞書に登録されている単語（登録語）だけを認識する構成となっている [1][2][3]. しかしながら音声認識システムの利用者が実際に音声認識を実行しようとすると，システムにどのような単語が登録されているのか知らないため登録されていない単語（未登録語）を含んだ発声をしてしまう場合がある．この場合音声認識システムは，本来未登録語である単語に対して登録語の中の何らかの他の単語に置き換えて認識結果を出力するため，しばしば全体の認識性能を低下させてしまう．また一方で実際の音声認識システムは辞書の大きさに対して制限があり，登録単語数を無限に増やすことはできない．このことからも音声認識システムでは，入力された発声の中に未登録語が含まれる可能性を常に考慮する必要がある．そこで我々は，未登録語を含んだ発声がなされた場合にも対処できるようなより頑健な音声認識方式の構築を目指している．

これまでに未登録語を含んだ発声を処理するための試みとしては，以下の方法が提案されている．

- 辞書に登録された単語の認識と並行して，未登録語として任意の音素連鎖を音素モデルを用いて認識する (本稿では，これを音素タイプライタ方式とよぶ) 方法 [4][5][6].
- 特定の品詞 (例えば固有名詞) に対して，辞書に登録された単語の単語モデルとともに未登録語用の単語モデルを駆動する方法 [7].
- 未登録語の対象を間投詞や不要語に限定し，対象となる単語のみから学習して作成した garbage HMM を用いる方法 [8][9].

これらの方法は，登録語のスコアと未登録語のスコアを同時に計算し，それぞれのスコアを比較することにより未登録語の検出を行うものである．音素タイプライタ方式は，未登録語の検出と同時にその音素系列を同定できるという利点はあるが，あらゆる音素の組合せを計算するため処理量が非常に大きいという問題がある．未登録語用の単語モデルを駆動する方法および garbage HMM を用いる方法は，未登録語の音素系列を知ることはできないが未登録語を単一のモデルで表現できるため音素タイプライタより処理量の増加が少ない．しかし一方でこの未登録語用モデルが，一つのモデルでさまざまな音素系列を表す荒いモデルであるため検出性能の低下が避けられない．このように一般に未登録語を扱うための処理を追加すると，処理量の増大および認識性能の大きな低下という問題が起る．

このような問題を踏まえて本稿では，処理量の増加を抑えつつ登録語に対する認識性能をできるだけ損なわずに効果的に未登録語を検出することのできる手法を検討する．ここでは，音素を数種類のクラスタに分類し，クラスタ毎に作成した音素クラスタモデルを用いた未登録語検出法を提案する．

2 未登録語の出現傾向

未登録語の出現傾向を調べるために，以下の調査を行った．国際会議予約に関する参加者と事務局の対話テキストデータ (650 会話; 総単語数約 3.3×10^5; 語彙数約 1.0×10^4) を用いて，辞書の語彙数を変えた場合の 1 会話中に含まれる単語のカバー率を調べた．なおこのカバー率は，以下の手順で計算した．

1. 使用するデータの量 n を会話単位 ($D_1,\cdots,D_n 1 \leq n \leq 650$) で決定する

2. 使用データの中から評価用データとして1会話 D_i を取り出す

3. 残った使用データ $(D_1, \cdots, D_{i-1}, D_{i+1}, \cdots, D_n)$ 中の出現単語から辞書を作成する

4. 作成した辞書を用いて，評価用データ D_i 中の単語のカバー率を計算する

5. 使用データ中の全会話 $(1 \leq j \leq n)$ について (2) ～ (4) を繰り返し，カバー率の平均を求める

6. (1) に戻り，データ量 n を増加させる

会話単位でデータ量を増やした場合の辞書の語彙数の増加を図0.1に，またその時の単語のカバー率を図0.2に示す．図0.1から，単一タスクである本テキストデータは，会話数450で語

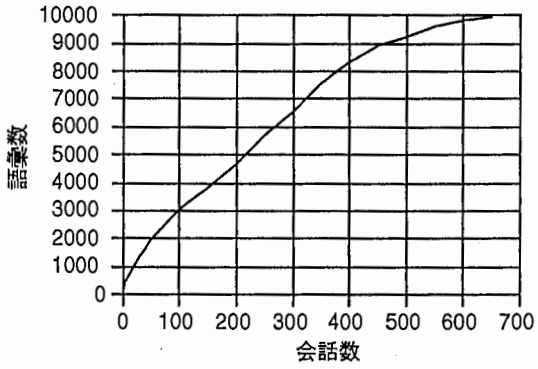

図 0.1: 会話数と語彙数の関係

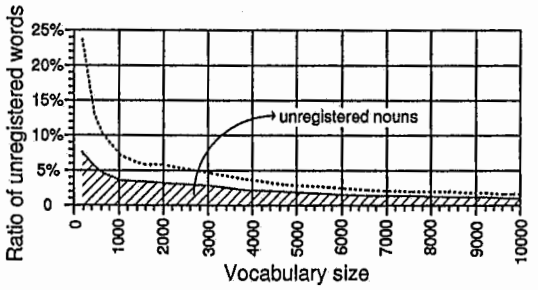

図 0.2: 辞書の語彙数に対する1会話中の単語の未登録語率

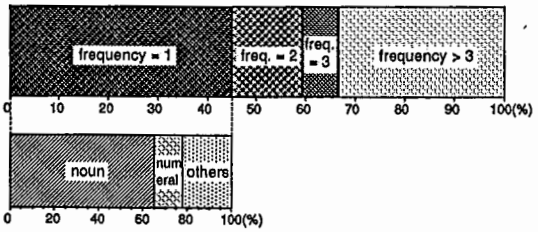

図 0.3: 単語の出現頻度の傾向

彙数約9,000に達し，語彙数の増加の割合が減少することが分かる．また図0.2では，語彙数7,000以上では未登録語となる割合が約2%でほぼ一定である．さらに，評価用データ中で未

登録語とみなされたすべての単語を30種類の品詞に分類してその内訳を調べると，辞書の語彙数が約3,000以上になると約60%が名詞(普通名詞，固有名詞)となることが分かった．この結果から，十分大きな語彙を持つ辞書を用意しても2%程度の単語が常に未登録語となる可能性が高いと考えられる．また今回使用したテキストデータでは，未登録語になる単語は品詞別に見て名詞が最も多いことが分かる．

一方，認識システムの辞書の作成について考えると，出現頻度の高い単語は辞書に登録されやすいが，出現頻度の低い単語は登録されにくいと想定できる．そこで，上記と同じテキストデータを用いて，単語の出現頻度の傾向を調べた．この結果を図0.3に示す．

その結果，未登録語になりやすいと考えられる出現頻度1の単語が全異り語彙の44.9%であった．またその内の65.2%が名詞(普通名詞，固有名詞)であることが分かった．以上辞書作成の点からも，未登録語になる単語は品詞別に見て名詞が最も多いと考えられる．本稿では，未登録語の出現傾向の結果を基に最も未登録語となりやすいと考えられる名詞に着目し，音声認識の文法中の名詞部分に対して未登録語の生成を許す規則を追加したものを用いることとした．

3 連続音声認識における未登録語処理

未登録語は，入力音声中に必ず存在するわけではないが，どこに存在するのか分からないため常時検出処理をする必要がある．そこで本稿では，少ない処理量で未登録語を検出し，未登録語が存在する場合のみ未登録語の存在する区間の音素系列を詳細に認識する以下の2段階のステップで未登録語を処理する音声認識システムを考えることにした．この音声認識システムを図0.4に示す．

Step 1: 登録語を音素モデルを用いて認識し，未登録語を音素クラスタモデルを用いることにより少ない処理量で検出

Step 2: 検出した未登録語を含む候補に対して，より精度の良い音素モデルを用いて音素系列を認識

本稿では，主にStep1の未登録語検出の処理について述べる．

4 音響モデル

提案する未登録語検出法では，表0.1に示す音響モデルを使用した．

本稿では，音素を数種類のクラスタに分類し，クラスタ毎に作成した音素クラスタモデルを用いた未登録語検出法を提案する．音素クラスタモデルは，さまざまなクラスタ構成をもつモデルが考えられる．今回は以下の7種類の音素クラスタモデルを作成し，それぞれの性能を比較検討した．なお音素クラスタモデルは，クラスタの構成に関わらず総混合数を一定の500とし，総状態数をクラスタ数nの時$4n$，各クラスタの混合数をそれぞれのクラスタが包含する音素数mに応じて$5m$混合とした．

(A-9) 日本語の音節構造を考慮して9個に分類した9クラスタモデル

日本語の音節は常に母音を含んだ構造をもつことから，5母音をそれぞれ単独のクラスタとする．母音は子音と比べて大量の学習データを持つため比較的信頼性が高いモデルが作成できると考えられる．また，子音の中でも拗音，促音，撥音は，音節構造からみて特殊な扱いが必要なためそれぞれ単独のクラスタとする．他の子音(母音との連結で音節を構成)は，すべてまとめて1つのクラスタとする．

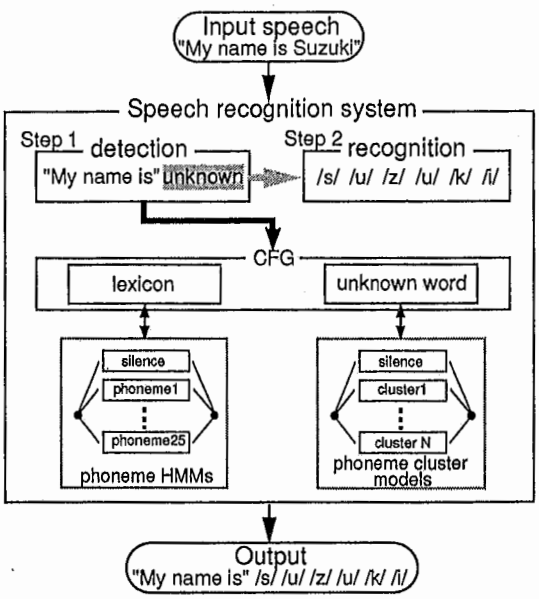

図 0.4: 未登録語処理を持つ音声認識システム

表 0.1: 使用した音響モデル

コンテキスト非依存モデル【Step1】
登録語
音素 HMM： 25 音素,100 状態,500 混合 (各音素モデル:4 状態, 各状態:5 混合) 無音 HMM： 4 状態,20 混合
未登録語
音素クラスタモデル： n クラスタ,$4n$ 状態,500 混合 (各クラスタモデル:4 状態, 各状態:$5m$ 混合) (m：クラスタ内の音素数)

(B-1) 全音素を 1 つにまとめた 1 クラスタモデル

25 音素をすべてまとめて 1 つのクラスタとする．(このクラスタモデルは，未登録語が単一のシンボルの系列として表される．このため音素 N-gram のような統計的言語モデルを制約として使うことができない．)

(C) 音響的な特徴を基に自動分割したクラスタモデル

音響的な特徴の差異をクラスタに反映させるために，全音素を 1 状態の HMM にまとめたモデルを初期状態として SSS(Successive State Splitting) により自動的にコンテキスト方向にのみ分割して [10]，その分割過程から以下 4 種類のクラスタ構成のモデルを作成した．

(C-3) 3 クラスタモデル

(C-5) 5 クラスタモデル

(C-9) 9 クラスタモデル

(C-11) 11 クラスタモデル

(D-25) 全音素が独立した 25 クラスタモデル
このモデルは，コンテキスト非依存型音素モデル (25 音素がそれぞれ独立したクラスタ) である．このモデルを駆動する未登録語処理は，従来法の音素タイプライタ方式 [4] に相当する．

作成したクラスタモデルのクラスタ構成を表 0.2に示す．但し表中 / はクラスタの区切り，q は促音，ng は撥音を表す．

表 0.2: 作成したクラスタモデルのクラスタ構成

(A-9)	/a/i/u/e/o/j/q/ng/j,k,zh,z,d, m,g,ch,r,sh,ts,s,b,t,w,n,p,h/
(B-1)	/a,i,u,e,o,q,ng,j,k,zh,z,d,m, g,ch,r,sh,ts,s,b,t,w,n,p,h/
(C-3)	/ng/e/a,o,w,i,u,q,j,k,zh,z,d, m,g,ch,r,sh,ts,s,b,t,n,p,h/
(C-5)	/ng/e/o/i,u,q/a,w,j,k,zh,z,d, m,g,ch,r,sh,ts,s,b,t,n,p,h/
(C-9)	/ng/e/o/i,u/a,sh,ts,s/j,zh,r,w/ k,ch,t,p,h,g/m,n,z,d,b,/q/
(C-11)	/ng/e/o/i/u/a/sh,ts,s/j,zh,r,w/ k,ch,t,p,h,g/m,n,z,d,b,/q/

5 音素クラスタモデルを用いた未登録語処理

提案する未登録語検出法の性能を評価するために，フレーム同期型 HMM-LR[12] を用いて未登録語検出実験を行った．実験条件を表 0.3に示す．

表 0.3: 実験条件

分析条件
サンプリング周波数 12kHz 20ms ハミング窓，フレーム周期 5ms
使用パラメータ
log power + 16 次 LPC-Cep + Δlog power + 16 次 ΔLPC-Cep

5.1 特定話者の未登録語検出 (Step1) の実験

未登録語の検出実験を行うために，以下の方法で未登録語を含んだ文を意図的に作り出し特定話者の文認識実験を行う．ここでは全認識データ 57 文を処理することの可能な文法を基にして各入力文に対して，

1. 文法の名詞部分に未登録語の生成を許す規則を追加

2. 入力文中 (正解単語系列) の名詞の有無を確認

3. 名詞があれば，単語辞書から該当する名詞のみ削除

4. 入力文中の名詞のみ削除された辞書を用いて認識

を行う．なお，入力文中の未登録語の数の制約は行っていない．

特定話者の未登録語検出の実験条件を表 0.4に示す．

表 0.4: 未登録語検出の実験条件

学習条件	
話者	男性 1 名 (MAU)
学習データ	重要語 2620 単語
音響モデル	コンテキスト非依存型音素 HMM, 音素クラスタモデル
認識条件	
話者	男性 1 名 (MAU)
認識データ	国際会議予約 57 文 (363 単語)
使用文法 (CFG)	国際会議予約 (語彙数:454) + 名詞部分に未登録語生成規則
ビーム幅	3000

ここで，作成した各クラスタモデルを用いて未登録語検出能力の比較実験を行なった．未登録語処理では，任意の連鎖を許す制約の弱い未登録語に対して何らかの閾値やペナルティを与えて，登録語のスコアと比較する方法が有効であることが知られている [5][6]．そこで本実験では，認識スコア S_{total} を次の式で求めた．

$$S_{total} = \sum_{t=1}^{n} S_t \tag{0.1}$$

但し，

【登録語のスコア】

$S_t = L_h$

【未登録語のスコア】

$S_t = \omega_2 * (L_{cl} + \omega_1 * log(P_{4-gram}))$

ここで，n はその時点でのフレーム数，L_{ph} は登録語の認識に使用した音素 HMM の音響尤度，L_{cl} は未登録語の検出に使用した音素クラスタモデルの音響尤度，P_{4-gram} は未登録語のクラスタの 4-gram の確率，ω_1 はクラスタ 4-gram に対する重み，ω_2 は未登録語のスコアに対する重みである．このクラスタ 4-gram は，今回の実験が名詞部分のみ未登録語を許す文法を使用していることから，国際会議予約に関するテキストデータ中から名詞のみ (全名詞数 58,896; 名詞種類数 5,072) を用いて学習した．以下，実験の評価は式 (0.2) による単語 accuracy で行った．

$$\text{単語 } accuracy = \frac{M - I - D - S}{M} \times 100\ (\%) \tag{0.2}$$

ここで，M は評価文の総単語数，I は挿入誤り，D は削除誤り，S は認識誤りの数を表す．但し，未登録語区間に未登録語系列が現れた場合は，正解単語として集計する．

比較する音素クラスタモデルは，異るクラスタ構成を持つモデルであるためそれぞれ計算される音響尤度は異る．また同様に，各クラスタで学習された 4-gram の確率値 P_{4-gram} も異る (表 0.5参照)．つまり未登録語のスコアに使用する重み ω_1,ω_2 は，音素クラスタモデル毎に最適値が異ると考えられる [11].

表 0.5: 学習したクラスタ 4-gram のパラーメータ数

クラスタ構成	パラーメータ数
(A-9)	1,107
(C-3)	90
(C-5)	437
(C-9)	2,243
(C-11)	3,414
(D-25)	9,183

本実験では，それぞれのクラスタモデルに対して適した重みを与えるために，2 つの重み ω_1,ω_2 についてそれぞれのクラスタモデルで以下 2 つの文単位の誤りを計算し，共に小さい値になる値を求め決定した (表 0.6の () 内参照).

- 未登録語を含まない文に対して未登録語を検出（未登録語の湧きだし誤り）
- 未登録語を含む文に対して未登録語を未検出（未登録語の削除誤り）

表 0.6に未登録語検出実験の結果および各クラスタモデルの処理時間の比を示す．　なおこ

表 0.6: 音素クラスタモデルを使用した未登録語検出実験

使用クラスタモデル (ω_1,ω_2)	単語 accuracy	処理時間の比 (sec)
(A-9) (1.0,0.99)	74.9%	0.52
(B-1) (0,0.98)	48.8%	0.34
(C-3) (1.0,1.0)	66.4%	0.42
(C-5) (1.0,1.01)	71.1%	0.45
(C-9) (1.0,1.01)	72.7%	0.70
(C-11) (1.0,1.0)	72.5%	0.70
(D-25) (1.0,1.02)	77.1%	1
未登録語処理なし	56.7%	0.38

の認識システムで，全名詞が登録されている (未登録語なしの場合にあたる) 辞書を使用した場合の単語 accuracy は 89.8% である.

表 0.6から，(B-1) のクラスタの場合は未登録語処理なしの場合の単語 accuracy より低い．しかし他のクラスタモデルについては，未登録語のスコアに対してクラスタの 4-gram を用いることで未登録語処理なしの単語 accuracy より改善できることが分かる．このクラスタの 4-gram は，クラスタの構成によってパラメータ数が異なり，クラスタ数を少なくするとパラメータを削減する点では有利であるが，言語制約の効果は小さくなる.

表0.6の結果から，(D-25) のモデルが最も高い検出性能を持つものの処理時間が多いのに対して，(A-9) のモデルは 25 クラスタに近い検出性能を持ち処理時間も 52% に削減できることが分かる．このことから，日本語の音節構造によるクラスタモデルが未登録語検出に有効であると考えられる．

5.2 不特定話者の未登録語検出 (Step1) の実験

次に，話者適応を行った音響モデルを用いて男女各 3 名についても同様の実験を行った．この結果を表 0.7に示す．音響モデルは，男性話者 (MHT) のモデルを男性 3 名に，女性話者 (FYM) のモデルを女性 3 名に，それぞれ 50 単語の発声を用いて移動ベクトル場平滑化方式 (VFS) [13] により話者適応を施したものを用いた．(音素クラスタモデルは，1 つのクラスタが複数音素から構成されることがあるため音素バランスのとれたデータを用いた話者適応を行う必要がある．) また，未登録語のスコアの重みは特定話者の場合と同様に誤りを計算し，共に小さい値になる値とした．クラスタ構造の違いによる影響を比較するためクラスタモデルには，クラスタ数の等しい (A-9) および (C-9)，また音素タイプライタ方式との比較のため (D-25) を使って実験を行った．

表 0.7から，50 単語の話者適応では未登録語検出能力としてはかなり低いもののクラスタ数が 9 でクラスタ構成の違う 2 つのモデル (A-9) および (C-9) は，話者適応してもほぼ同等の単語 accuracy を達成している．これは，クラスタ数が 9 あれば SSS により決められたクラスタであっても日本語の音節構造に近い記述ができるためだと考えられる．また，(D-25) を使った音素タイプライタ方式は，音素クラスタモデルと比較して単語 accuracy が低下している．このことからも日本語の音節構造によるクラスタモデルは，話者適応を行った不特定話者の未登録語検出においても有効であると考えられる．

ここで音素クラスタモデルの適切なクラスタの設計を考えてみると，SSS を用いて決定したクラスタは話者に依存した話者固有のクラスタ構成であるのに対して，日本語の音節構造に従ったクラスタは話者に依存しない．このことから SSS により決定したクラスタは，話者適応を行うと適応話者本来のクラスタ構成と異るクラスタ構成を使用することになり，認識性能の低下の危険がある．さらに，クラスタの N-gram の学習の点においても話者に依存しない日本語の音節構造に従って決めたクラスタは，話者毎に学習する必要がないため有利だと考えられる．

5.3 検出した未登録語の音素系列認識 (Step2) の実験

本稿で提案する特定話者の未登録語検出の結果を用いて，検出された未登録語に対して音素系列認識が可能であることを確かめるために，より精度の良い音素モデルを使い以下の実験を行った．Step1(未登録語検出) の結果には，前節で最も未登録語検出に有効とみなした (A-

表 0.7: 話者適応を施したクラスタモデルを用いた男女各 3 名の単語 accuracy (%)

標準話者	MHT			FYM			6 話者の	6 話者の
適応・評価話者	MAU	MIK	MTK	FAK	FAS	FNY	average	処理時間の比
(A-9)	64.7	55.9	57.0	69.1	55.1	48.2	58.3	0.65
(C-9)	67.5	53.2	61.4	69.4	53.7	45.5	58.5	0.76
(D-25)	68.3	51.5	58.7	58.1	53.7	51.2	56.9	1

9) のモデルの結果を用いた．(A-9) のモデルによる未登録語検出の結果の内 1 位に未登録語を検出した文は，31 文 (未登録語を含む 33 文中から 30 文，未登録語を含まない 24 文中から 1 文) であった．Step2 の音素モデルにはコンテキスト依存型音素 HMM(HMnet:400 状態, 各状態 3 混合) を用い，認識スコア S_{total} は未登録語の検出と同様に式 (0.1) を用いた．音素系列認識実験の評価は，登録語と未登録語を含んだ文全体の音素 accuracy で行った．ここでは未登録語の検出実験の結果を基にして，

1. 未登録語検出の結果の第 1 位の候補に対して未登録語の有無を確認
2. 未登録語があれば，Step1 の上位 10 位までの候補をそのまま認識候補とする
3. 候補内の未登録語部分を (音素連鎖の制約を使った) 音節連鎖の規則に変更
4. 作成された文法を用いて文全体を認識

の一連の処理を行う．

以下にこの実験で用いた文法の例 (候補が 2 個の場合) を示す．

【未登録語の検出結果】

- 名前は ” 未登録語” です
- 名前 ” 未登録語” です

【実験に使用する文法】

- 文 → 名前は < 音節連鎖規則 > です
- 文 → 名前 < 音節連鎖規則 > です

ここで，音節連鎖規則は Step1 で用いた規則と同様のものである．

未登録語の音素系列認識の実験条件を表 0.8に示す．

表 0.8: 未登録語の音素系列認識の実験条件

学習条件	
話者	男性 1 名 (MAU)
学習データ	重要語 5240 単語
音響モデル	コンテキスト依存型音素 HMM (HMnet)
認識条件	
話者	男性 1 名 (MAU)
認識データ	未登録語検出の結果で 1 位の中に未登録語を検出した文
使用文法 (CFG)	各認識対象文の未登録語検出結果上位 10 候補の文 + 未登録語部分に音節連鎖規則
ビーム幅	3000

また，Step2 では音素系列の長さに対する自由度を小さくするために，音節数の制約を入れることとした．方法としては，未登録語検出で使用した (A-9) のモデルが音節構造を考慮し

て決められたモデルであることを利用して，検出された未登録語の音節数を基準に音節の連鎖数を制約した．2段階で認識する本方法に対してこの制約を使用した場合，また1段階で認識する音素タイプライタ方式(Step1で用いた文法と同じ文法を使用)にコンテキスト非依存型音素HMM(D-25)およびコンテキスト依存型音素HMM(HMnet)を用いた場合の前述の31文に対する結果を表0.9に示す．

表0.9: 提案法および従来法による音素系列認識の実験

	使用音素HMM (ω_1,ω_2)	音素 accuracy
2段階	コンテキスト依存 (1.0,1.0)	76.7%
1段階	コンテキスト非依存 (1.0,1.02)	79.3%
	コンテキスト依存 (1.0,1.0)	77.7%

表0.9から，(D-25)のクラスタモデルを用いた方法が音素accuracyが最も高いものの，2段階の処理は認識性能の劣化は3%以内に留まっている．

次に認識対象の国際会議予約57文を対象に，提案法および従来法で未登録語処理を行った場合の比較を行った．但し2段階法の音素accuracyは，57文の内未登録語を検出しなかった26文は未登録語検出の結果を使い，未登録語を検出した31文は音素系列認識の結果を使って計算した．2段階法の音素系列認識の処理時間は，音素系列認識を行った31文の結果のみを使って計算した．また，1段階のHMnetを使用する方法の単語accuracyおよび音素accuracyは，57文中2文の認識がビーム幅の不足によりできないため，この2文はすべて削除誤りとして計算した．この結果を表0.10に示す．

表0.10: 提案法および従来法による未登録語処理実験

2段階		
	未登録語検出	音素系列認識
使用モデル (ω_1,ω_2)	(A-9) (1.0,0.99)	HMnet (1.0,1.0)
単語 accuracy	74.9%	–
音素 accuracy	–	80.5%
処理時間の比	0.52	0.57
1段階(コンテキスト非依存型音素モデル)		
使用モデル (ω_1,ω_2)	(D-25) (1.0,1.02)	
単語 accuracy	77.1%	
音素 accuracy	82.3%	
処理時間の比	1	
1段階(コンテキスト依存型音素モデル)		
使用モデル (ω_1,ω_2)	HMnet (1.0,1.0)	
単語 accuracy	70.2%	
音素 accuracy	80.0%	
処理時間の比	0.87	

表0.10から，単語accuracy，音素accuracy共に25クラスタモデル(コンテキスト非依存型音素モデル)を用いた1段階の方法が最も高い．一方認識処理時間では，HMnet(コンテキ

スト依存型音素モデル) を用いた 1 段階の方法が最も短い．2 段階の方法では，提案した検出法の処理については高速だが全体では 1 段階の方法よりも遅い．しかし，今回の 2 段階法の音素系列認識の実験では文全体 (登録語も含む) を再度 HMnet を用いて認識する方法を用いているが，未登録語検出の時点で未登録語の存在する区間を参照し，その区間でのみ音素系列の認識を行えば大幅に処理時間を短縮することが可能である．今後は，さらに Step2 の処理の高速化を検討する必要がある．また，ビーム幅に関しては実験では同じビーム幅を用いた．1 段階のコンテキスト依存型音素モデルの認識性能が低い原因は，このビーム幅が充分に確保できない点にある．しかし 2 段階法の未登録語検出および音素系列認識は共に，1 段階の方法に比べて生成する仮説数が少ないことからビーム幅をさらに削減することが可能である．このことから，2 段階法はさらに全体の処理時間の削減を期待できると考えられる．

6　まとめ

本稿では未登録語を扱うことの可能な音声認識システムの構築を目指し，音素クラスタモデルを用いた未登録語の検出法について提案した．未登録語検出実験では，数種類の音素クラスタモデルを用いることにより，従来の音素モデルを用いた音素タイプライタ方法と比較して，処理量を約 52% に削減しながらほぼ同等の単語 accuracy (74.9%) を達成することができた．このことから，音素クラスタモデルを用いる方法が処理量を抑えた未登録語検出に有効であることが分かった．またこの時，未登録語のスコアに対してクラスタ N-gram の確率をペナルティとして使用することが，有効であることが分かった．次に未登録語検出と検出された未登録語の音素系列の認識を 2 段階で行う方法により，1 段階の方法である音素タイプライタ方法とほぼ同等の音素 accuracy (76.7%) を達成することが確認できた．また 2 段階法の音素系列認識は，1 段階目の結果を用いて処理区間を制限することや各段階のビーム幅をさらに小さくすることが可能と考えられ，処理量の削減が期待できる．

今後は，少ない処理量で効率良く未登録語の検出を可能にするより頑健な音声認識方式の構築を目指す上で，音素のクラスタ構成・未登録語スコアのペナルティ・最適なビーム幅の設定を検討する必要がある．また，2 段階目の処理の高速化についてもさらに検討する必要がある．

謝辞

研究の機会を与えて頂いた ATR 音声翻訳通信研究所 山崎泰弘社長，匂坂芳典室長に感謝いたします．また，熱心な御討論と有益な御助言を頂いた研究室の方々に感謝いたします．

参考文献

[1] M. Bates, R. Bobrow, P. Fung, R. Ingria, F. Kubala, J. Makhoul, L. Nguyen, R. Schwarts, D. Stallard, “The BBN/HARC Spoken Language Understanding System”, Proc. ICASSP93, pp.111-114, 1993.

[2] H. Ney, V. Steinbiss, R. Haeb-umvach, B. -H. Tran, and U. Essen, “An Overview of the Philips Research System for Large Vocabulary Continuous Speech Recognition”, International Journal of Pattern Recognition and Artificial Intelligence, Vol.8, No.1, pp.33-70, 1994.

[3] P.C. Woodland, J.J. Odell , V. Valtchev and S.J. Young, “Large Vocabulary Continuous Speech Recognition Using HTK”, Proc. ICASSP94, pp.125-128, 1994.

[4] K. Kita, T. Ehara, and T. Morimoto, "Processing unknown words in continuous speech recognition", IEICE Trans. Vol.E74, No7, pp.1811-1816, 1991.

[5] 渡辺隆夫, 塚田聡, "音節認識を用いたゆう度補正による未知発話のリジェクション", 電子情報通信学会論文誌, Vol.J75-D-II, No12, pp.2002-2009, 1992.

[6] 伊藤克亘, 速水悟, 田中穂積, "連続音声認識における未知語の扱い", 信学技報, SP91-96, pp.41-47, 1991.

[7] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system", Proc. ICASSP91, pp.305-308, 1991.

[8] 井ノ上直己, 武田一哉, 山本誠一, "ガーベジ HMM を用いた自由発話文中の不要語処理手法", 電子情報通信学会論文誌, Vol.J77-A, No2, pp.215-222, 1994.

[9] 山田雅章, 伊藤史郎, 酒井桂一, 小森康弘, 大洞恭則, 藤田稔, "音声対話 CD-ROM 情報検索システム - 対話中における未知語部分の再評価アルゴリズム -", 信学技報, SP93-21, pp.57-64, 1993.

[10] J. Takami, and S. Sagayama, "A Successive State Splitting Algorithm for Efficient Allophone Modeling", Proc. ICASSP92, pp.573-576, 1992.

[11] H. Sakamoto, S. Matsunaga, "Detection of Unknown Words using Garbage Cluster Models for Continuous Speech Recognition", Proc. EUROSPEECH'95, THpm1D.3, pp.2103-2106, 1995.

[12] T. Shimizu, S. Monzen, S. Matsunaga, and H. Singer, "Time-Synchronous Continuous Speech Recognizer Driven by a Context-Free Grammar", Proc. ICASSP95, pp. 584-587, 1995.

[13] 大倉計美, 杉山雅英, 嵯峨山茂樹, "混合連続分布 HMM を用いた移動ベクトル場平滑化話者適応方式", 信学技報, SP92-16, pp.23-28, 1992.