2005－SLP－57（20）
2005／7／16

未知固有表現を含む音声の認識

冨田 達彦[a]　沖本 純幸[b]　山本 博史[c]　匂坂 芳典[d]

a)c)d) 早稲田大学 国際情報通信研究科 〒 169-0072 東京都新宿区大久保 3-14-9
b) 松下電器産業（株） 先端技術研究所 〒 619-0237 京都府相楽郡精華町光台 3-4
c) ATR 音声言語コミュニケーション研究所 〒 619-0288 京都府相楽郡精華町光台 2-2-2

E-mail: a) 61403675@suou.waseda.jp, b) okimoto.yosh@jp.panasonic.com,
c) yama@slt.atr.co.jp, d) sagisaka@giti.waseda.ac.jp

要旨 任意の未登録語を含む音声認識を目指して、複数単語列からなる未登録固有表現を含む音声の認識実験を行なった。TV 番組検索タスクに対し、通常の単語クラス N-gram モデルを上層のモデルに、下層には未登録固有表現（TV 番組名）に対する単語連接制約モデル、未登録語（人名）に対する単語構造モデルの 2 つを作成した。音声認識実験により、新たに加えた下層の単語連接制約モデルの性能と、2 つの未登録語モデル間の交互作用を調べた。実験の結果、未登録表現箇所はほぼ正しく同定され、全未登録語・表現を辞書登録した性能上限モデルに迫る単語認識率を得た。また、未登録表現と異種の未登録語間の混在や誤認識は見られず、階層化言語モデルの単語連接への拡張可能性が判明した。

Speech recognition of unregistered expressions

Tatsuhiko Tomita[a]　Yoshiyuki Okimoto[b]　Hirofumi Yamamoto[c]　Yoshinori Sagisaka[d]

a)c)d) Global Information and Telecommunication Institute, Waseda Univ.
b)Advanced Technology Research Laboratories, Matsushita Electric Industrial Co., Ltd.
c) ATR Spoken Language Translation Research Laboratories

Abstract Aiming at speech recognition with arbitrary OOV expressions, speech recognition experiments were carried out using speech with OOV expressions consisting of multiple words. For a TV program retrieval task, a hierarchical language model was newly composed using conventional word class N-grams as an upper layered model and two lower layered models consisting of word N-grams for an OOV expressions (TV program names) and a statistical phonotactic word-structure model for OOV words of another class (personal name). Speech recognition experiment results showed reasonable performance of word N-grams for OOV expressions and no serious interference between two OOV models, which confirms the availability of a hierarchical OOV model with word-level statistics.

1　はじめに

自然な発話の中では、商品名やものの名前等、単語の組合せからなる未知の固有表現がよく用いられる。現在の音声認識の枠組みでは、これら登録辞書外の単語を含む発話への考慮は十分にはなされておらず、それらの完全な認識は難しい。この問題に対処するため、これまでに人名、地名といった単純な未登録語に対してモデル化が検討されている [1][2][3]。これら単一の未登録語に対しては、従来の単語クラス N-gram に加えて、未登録語彙の統計的な音韻連接制約 (phonotactics) を与える単語内モデルを合わせて使用した階層的言語モデルの有効性が確認されている [2]。階層化言語モデルは単一の単語クラスの未登録語のみならず、複数の単語クラスに対しても同様に有効であることも示されている [3]。都市名と人名の未登録語を含んだ音声の認識実験では、両単語クラス間の混同は小さく、独立した異種の言語統計量を一つの言語モデルとして実現している。

単一の未登録語に対しては、上述した階層化言語モデルの有用性が期待できるが、すべての未登録表現への扱いとしては不十分である。実際の発話では番組のタイトルや本の題名のように、複数の単語のまとまりが一つの未登録表現を構成する場合が多くみられ、これらは単一の未登録語のモデルとして取り扱うには無理があると思われる。本稿では、この問題に対処するため、未登録表現を構成する単語による統計的制約を下層のモデルに適用することを考えた。

以下、次章では本稿で扱う未登録固有表現について述べ、本検討の目的を示す。第 3 章では、複数の未登録語クラスに対処するために利用した階層化言語モデルについて説明し、第 4 章では音声認識実験による評価と考察について述べる。最後に、第 5 章でまとめと今後の課題について触れる。

2　未登録固有表現の認識

未登録語を含む表現としては種々のものが考えられるが、ここでは単一の単語のみならず、いくつかの単語がまとまった一つの表現 (named entity) としてみなせる単語連接まで拡張する。商品名や会社名といったものは、これまでに進められてきた人名、地名といった単純な未登録語に対するモデル化を援用することによる解決が期待できるが、複数の単語に対しては対処が必要であると考えられる。書名や映画のタイトル等のように、用いられる単語そのものは種々のものが有り得て、その箇所全体が一つの named entity として用いられるものの例は多い。人名、地名の認識でそれらの音韻認識が不十分でも単位の同定が有用であるのと同様に、構成単語の完全な認識が不十分でも単語の連接を一つの named entity として同定することは有用である。とりわけ、音声認識で未登録固有表現が引き起こす誤認識の防止、named entity 箇所の認識結果の利用等に役立つことが大きく期待される。

3　階層化言語モデル

図 1 に示すように、階層化言語モデルは 2 つの層からできている。上位の層は従来の単語クラス N-gram であり、通常の言語モデルとして用いられている部分である。階層化言語モデルでは、さらに下位の層としてクラス依存サブワード N-gram 群を備えている。それぞれのサブワード N-gram は上位層の単語クラスと対応している。これらは、対応するクラス特有の言語統計量によりクラス内の未登録語を受理する。階層化言語モデルでは、これらの N-gram モデル群を、その独立性を保ったまま、一つの言語モデルとして統合する。

下位層をなすそれぞれのサブワード N-gram は、対応する単語クラスについての一般的な単語コーパス（例えば、日本人名データベース）を用いて学習し、そのクラスの単語をサブワード列で作成する。この際、従来未登録語認識のターゲットとされていた日本人名などの単一な単語に対しては、統計的な音韻連接制約を用いた単位の有効性が示されている [3]。これは、形態素等の構成要素の使用特性に依存するためと考えられ、人名の場合以下のような特徴となって表れている。

- 長さに関する特徴
  3 ないし 4 モーラが一般的である
- 並びに関する特徴
  ナカ、ヤマなど高頻度の単位が存在する

これに対し、本稿で対処とする複数の単語からなる未登録固有表現では、上記のような読みに関する特徴があるとは考えにくく、音韻連接制約だけでは、認識が困難と考えられる。そこで、図 1 のように、未登録固有表現クラスに対しては、下位の層におい

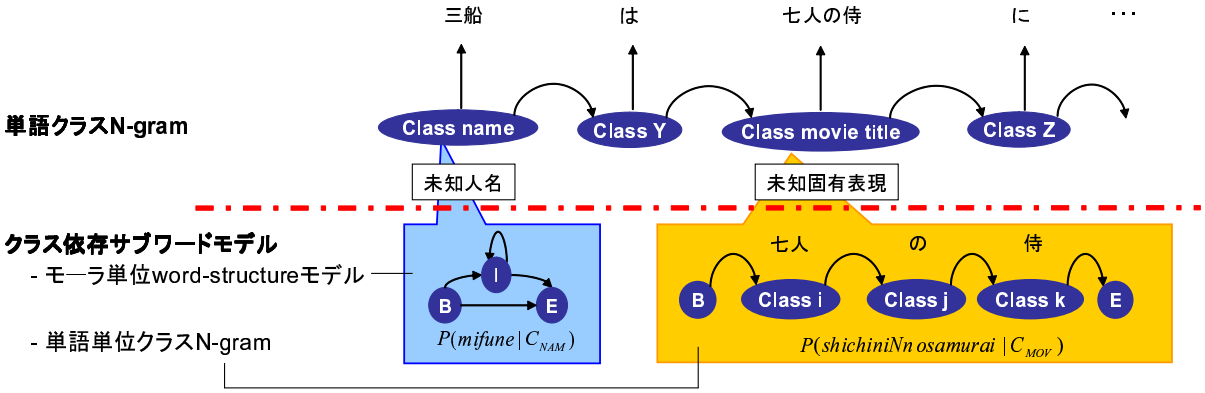

図 1　階層化言語モデルの構成（2 つの層からなり、それぞれが異なる統計的制約を与える）

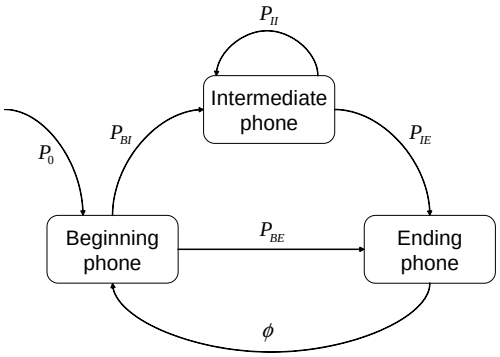

図 2　単語構成モデル（図中のアークは各サブワード bi-gram に対応）

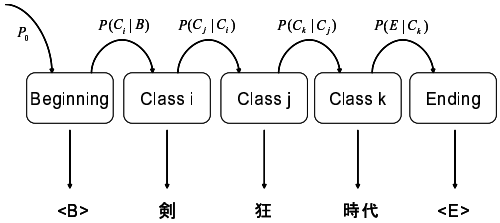

図 3　未登録固有表現の 2-gram モデル

て、上層と同じように単語単位の統計的制約を与えることを考えた。図 1 に示すように、本検討では未登録固有表現として TV 番組名クラスを、未登録語として人名クラスを考えた。以下、人名クラス、固有表現クラスの持つ下位層のモデリングについて、それぞれ記す。

3.1　単語構成モデル

ある単語クラスの特徴を捉えたサブワード N-gram の記述として、単語の生成特性を反映させた単語構成モデルが提案されている。このモデルは、人名クラスを少ない情報量で精度良くモデル化できることが示されている [3]。このモデルでは、図 2 に示すように、サブワード列の遷移を始端、中間、終端の状態で統計的に記述する。本稿でも、未登録人名クラスに対して、この単語構成モデルを適用する。

3.2　未登録固有表現モデル

複数の単語から構成される固有表現の統計的言語制約としては、図 3 に示すように単語クラス 2-gram を用いた。このモデルで、上位層と下位層において同じ単語が現れる場合があるが、下位層の単語は上位層のそれとは違うものとして学習している。例えば図中の「時代」という単語は、上位層と下位層においてそれぞれ異なった確率値を持つ。この理由は、固有表現の単語連接特性が、通常の文章とは異なることにほかならない。

4　音声認識実験

階層化言語モデルを利用して、人名・固有表現クラスが未登録である場合の音声認識実験を行なった。実験では提案するモデル（提案モデル）に加え、本

モデル化による認識性能の上限を与えると考えられるモデル（上限モデル）および従来の階層化言語モデルにより未登録語処理を施したモデル（従来モデル）の 3 種類を比較した。提案モデルは人名クラスをモーラ単位単語構成モデル、固有表現クラスを単語クラス 2-gram でモデリングした階層化言語モデルによって構成される。上限モデルは全ての単語を既知単語とした 1 層構造の言語モデルであり、未登録語が認識性能に影響を与えないため、ここで考慮する未登録語モデルの性能上限を示すものと考えられる。従来モデルは人名クラス、固有表現クラスを単語の代わりにモーラ単位を用いた単語構成モデルで対応する、従来の階層化言語モデルである。

4.1 認識対象

階層化言語モデルにおいて下位層を持つ 2 つの登録語クラスとして、次の 2 つを使用した。第 1 は人名クラス（日本人名クラス／外国人名クラス）、第 2 は未登録固有表現として TV 番組名クラスである。

学習データとして、人名テキストコーパスより固有姓名 39340 件、TV 番組名テキストコーパスより TV 番組 12185 件を用いた。

4.2 構築した言語モデル

3 種類の言語モデルは以下のように作成した。

- 提案モデル
  人名クラスを単語構成モデル、TV 番組名クラスを単語クラス 2-gram でモデリングした階層化言語モデル。下位層は上記のテキストコーパスより学習したもので、人名クラスは 3330 個のユニット（モーラおよびその結合したもの）、TV 番組名クラスは 8030 単語（連結単語を含む）を持つ。
- 上限モデル
  全ての単語を既知語とした、1 層構造からなる言語モデル。番組検索表現 3670 文において学習した単語クラス 2-gram[4] に、実験で用いた音声データに現れる人名、TV 番組名を加えたものである。提案モデル並びに従来モデルに用いる階層化言語モデルの上位層は、この上限モデル中の人名・TV 番組名をそれぞれの単語クラスに置換したものとなっている。
- 従来モデル
  人名クラス、TV 番組名クラスをそれぞれ単語構成モデルでモデリングした階層化言語モデル。下位層は上記のテキストコーパスより学習したもので、人名単語構成モデルは提案モデルと同じもの、TV 番組名クラスは 3255 個のユニットを持つ。

4.3 実験条件

上記の言語モデルを用いた音声認識実験を行なった。評価用音声データには、TV 番組検索表現 100 文を男女 7 人ずつ読み上げた音声、計 1400 文を用いた。このテストセットには、359 個の人名と 209 個の TV 番組名が含まれており、これらは上記の未登録語クラス学習用テキストコーパスからは除外している。音声の分析条件はサンプリング周波数 12kHz、フレーム長 20ms、フレーム周期 10ms とし、特徴量として MFCC12 次元＋そのデルタおよびデルタパワーの計 25 次元を用いた。音響モデルは 1400 状態 5 混合 HMnet model based on ML-SSS[5] 、デコーダは ATRASR[6][7] を用いた。ビーム幅、言語スコア重み、挿入ペナルティは、構築した言語モデルそれぞれにおいて単語正解精度が最大になるように選択した。

5 結果と考察

表 1 に、単語全体について測定した単語認識率を示す。ここで、階層化言語モデルに関しては、未登録語の認識結果はサブワード（モーラもしくは単語）で出力されるため、正解判定基準が必要である。ここでは、クラス情報（人名クラスか TV 番組名クラスか）が正しく、文中の正しい位置に出現している場合に正解とした。表 1 を見ると、従来モデルに比べて提案モデルの単語 correct、accuracy の値がそれぞれ 10 ポイント程度上昇しており、上限モデルに近接した数値をとっている。この結果から、固有表現をモーラ単位から単語単位でモデリングすることにより、未登録語を含む文章全体をより精度良く認識できることがわかる。

表 1　単語全体に関する単語認識率

言語モデル	単語 correct	単語 accuracy
提案モデル	97.07	94.91
上限モデル	98.60	97.97
従来モデル	87.12	83.72

表 2　未登録語に関する再現率・適合率

言語モデル	再現率		適合率	
	人名	番組名	人名	番組名
提案モデル	98.6	100	59.2	83.3
比較モデル	93.8	98.4	47.0	59.6

表 2 は、人名クラス・TV 番組名クラス、すなわち未登録語クラスの再現率・適合率を示している。再現率と適合率は次式で与えられる。

$$\text{再現率} = \frac{\text{認識結果中の正解未登録語数}}{\text{正解中の未登録語数}} \times 100$$

$$\text{適合率} = \frac{\text{認識結果中の正解未登録語数}}{\text{認識結果中の未登録語数}} \times 100$$

表 2 に示されるように、提案モデルも従来モデルも高い再現率を測定しており、特に、提案モデルではほぼ 100%に近い値をとっている。このことから、階層化言語モデルを用いることにより高い精度で未登録語クラスを認識できていることがわかる。次に適合率を見ると、かなり低めの値になっている。これは、未登録語クラスの湧き出し誤りの多さに起因する。しかし、従来モデルに比べて提案モデルでは、人名クラス 12.2ポイント、TV 番組名クラス 23.7ポイントの向上が見られた。この結果は、従来モデルは未登録語をモーラ単位で出力するため未登録語の無い箇所でも出力が強く制限されないのに対し、提案モデルの TV 番組名は単語単位で出力するため、言語制約が強く働き未登録語の湧き出しを防いでいることによると考えられる。

表 3 に未登録語クラスの音素認識率を示す。この値は、未登録語の場所を正しく同定した部分においての音素認識率である。提案モデルでは、従来モデルに比べ、音素 correct4.7ポイント、音素 accuracy6.5ポイントの向上が見られた。実験では不要な湧き出し誤りを防止する目的で適度の挿入ペナルティをかけているため、未登録部が長く続くことを制限する。このため、従来モデルでは未登録固有表現のモーラ列全てを出力することは困難となる。一方、提案モデルでは、単語単位のモデル化をすることにより未登録固有表現を構成するユニット数が減り、正しく未登録部が同定された結果、音素認識率が向上したと考えられる。

表 4 に、階層化言語モデルを用いた認識実験の際に、人名クラスと固有表現クラスが誤ったクラスに置き換わった個数とその置換率（confusion rate) を示した。これらの置換誤り数は、置換で生じた単語クラスごとに集計してある。置換率の導出の仕方は以下の通りである。

表 3　未登録語の音素認識率

言語モデル	音素 correct	音素 accuracy
提案モデル	82.78	82.01
従来モデル	78.08	75.51

表 4　未登録語クラスの confusion rate

言語モデル		認識対象		confusion rate
		人名	番組名	
提案モデル	人名	326	47	13.91
	番組名	32	163	
	その他	0	0	
従来モデル	人名	224	60	33.45
	番組名	130	145	
	その他	5	4	

$$confusion\ \ rate = \frac{\text{置換誤り数}}{\text{総クラス数}} \times 100$$

この結果によれば、従来モデルでは人名クラスを 135 個、番組名クラスを 64 個置換誤りしているのに対し、提案モデルでは人名クラスを 32 個、番組名クラスを 47 個置換誤りと、その数を大きく減少している。これより、提案モデルにより未登録語クラス同士の相互干渉を防いでいることがわかる。

表 5 に、番組名クラスの内容語同定率を示す。内容語同定率の値は、未登録固有表現を構成する単語のうち、名詞や動詞等の意味を持つ内容語をどれだけ同定できたを表わす。表 5 が示すように、提案モデルでは 60.42%の内容語を同定しているのに対し、従来モデルでは内容語を同定できていない。これは、従来モデルがモーラ単位で認識結果を出力するためである。この結果は、提案モデルが後段の意味処理等の言語処理に有用な情報を提供する可能性を示唆している。

表 5　番組名クラス内容語同定率

言語モデル	内容語 correct
提案モデル	60.42
従来モデル	-

6　まとめ

本稿では、タスクに依存しない一般的な言語モデルの構築を目指し、複数の単語からなる未登録固有表現（TV 番組クラス）と異種の未登録語（人名）を含んだ音声について階層化言語モデルを用いた認識実験を行った。階層化言語モデルの下位層において、未登録固有表現を単語単位でモデル化する未登録固有表現モデルを用いることで、従来のモデルを上回る認識率を得た。文章全体の単語認識率は、発話中の単語すべてを既知語とした場合の上限モデルに近接した認識率を示した。さらに、人名クラスと TV 番組クラスという 2 つの未登録語クラス間において、一方を他方に誤認識する割合を測定したところ、従来のモデルに比べて大幅に置換誤り数が減少した。これらの結果より、未登録固有表現を単語単位でモデル化することが有効であり、未登録表現・語モデル間の相互干渉も少ないことが判明した。

今後は認識率の向上を図るとともに、未登録語クラスの湧き出し誤り数をさらに削減し、タスクフリーな言語モデルの構築を目指したいと考えている。

謝辞

本研究は科学研究費基盤研究 (B)(2)-14380168「未登録語を含むタスク外発話を受理する音声認識方式の研究」、早稲田大学特定課題研究奨励費 2002B-038 の援助による。

参考文献

[1] I. Bazzi, R. Glass, “Modeling Out-Of-Vocabulary Words For Robust Speech Recognition,” Proc. ICSLP, Vol1. pp. 401-404, Beijing, 2000

[2] 谷垣 宏一, 山本 博史, 匂坂 芳典, “クラスに依存した語彙の確率的記述に基づく階層型言語モデル, ” 信学論, D-II Vol. J84-D-II, No.11, pp. 2371-2378, 2001

[3] 大西 茂彦, 小窪 浩明, 山本 博史, 匂坂 芳典, “大語彙連続音声認識における未知語の sub-word モデリング手法, ” 信学技報, SP2001-5, pp. 33-39, 2001

[4] 山本 博史, 匂坂 芳典, “接続の方向性を考慮した多重クラス複合 N-gram 言語モデル, “ 信学論, D-II Vol. J83-D-II pp. 2146-2151, 2000

[5] M. Ostendorf, H. Singer, “HMM topology design using maximum likelihood successive state splitting,” Computer Speech and Language, 11(1):17–41, 1997

[6] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga, Y. Sagisaka, “Spontaneous dialogue speech recognition using cross-word context constrained word graphs,” Proc. ICASSP 1996, pp. 17–41, 1996

[7] 伊藤玄，葦苅豊，實廣貴敏，中村哲，“音声認識統合環境 ATRASR の概要と評価報告, ” 音響講論,1-P-30, pp.221-222, 2004

[8] T. Hazen, I. Bazzi, “A COMPARISON AND COMBINATION OF METHODS FOR OOV WORD DETECTION AND WORD CONFIDENCE SCORING,” Proc. ICASSP, Salt Lake City, 2001

[9] T. Schaaf, “Detection of OOV Words Using Generalized Word Models and a Semantic Class Language model, ” Proc. EuroSpeech ,Scandinavia, Vol4. pp. 2581-2584, 2001

[10] 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, 1991