サブワード モデルを用いた未登録語認識の効率的探索手法

小窪浩明[†] 大西茂彦[†]
山本博史[†] 菊井玄一郎[†]

本論文では，未登録語を含む音声の認識を可能とするクラス依存サブワード モデルを効率的にデコードする手法として，サブワード ネットワークを用いたデコーダを提案した．提案したデコーダを従来の単語 N-gram ベースのデコーダと比較したところ，認識性能を劣化させることなく言語モデルのデータサイズを 1/40 にし，46%の処理量削減を実現した．また，提案したサブワード ネットワークの構造を応用し，日本人姓/名を対象としたモデル化に有効と思われる言語的特徴量について検討した．評価用音声データに出現する未登録語の認識実験の結果，サブワード bigram のみでモデル化した場合に比べ，モーラ長確率や単語終端位置での生起確率を特徴量に追加することで，未登録語の正解数が約 15%向上した．

Efficient Decoding Method for OOV Words Recognition with Subword Models

Hiroaki Kokubo,[†] Shigehiko Onishi,[†] Hirofumi Yamamoto[†] and Genichiro Kikui[†]

Class dependent subword models were found to be effective for recognizing OOV (out-of-vocabulary) words. This paper proposes a novel decoder that efficiently handles the models. Compared with previous decoder, the proposed method achieves language model size of 1/40, and 46% reduction in CPU time without any deterioration of performance. Then, using the structure of subword networks, we examine feature parameters of subword models, which are applied to Japanese family/personal name. The result of speech recognition for OOV words indicates that by using of additional characteristics (e.g., duration or occurrence probability in word-end), the number of correctly recognized OOV words was improved by about 15%.

1. はじめに

音声認識の大語彙化が進んでいるにもかかわらず，未登録語の問題は依然解決していない．特に，人名や地名などの固有名詞は語彙を増加してもそのすべてを網羅することは困難である．一方，固有名詞はタスク達成上重要な情報を多く含んでおり，固有名詞の未登録語認識は大きな技術課題の 1 つとして位置付けることができる．

未登録語を含む連続音声認識の先行研究としては，登録語を認識対象としたデコーダと並行して音素タイプライタを動作させ，未登録語の音素系列を推定する方法[1),2)]，未登録単語の構成単位をサブワードとして新たに単語辞書に登録しておく方法[3)~5)]などが提案されている．音素タイプライタを並行して動作させる前者の方法は，2 つのデコーダを動作させる必要があるため，処理量の観点から不利な場合も多く，甲斐ら[2)]は未登録語の探索を独立に行うことで処理の効率化を図っている．また，推定される未登録語区間の音響スコアは最尤音素系列のスコアが使われるため，語彙内の単語仮説と統合するにはペナルティなどのヒューリスティックスが必要となる．一方，後者のサブワードを単語辞書に登録させる方法では，デコーダにいっさい手を加えることなく実装が可能であるという利点はあるものの，サブワード間の探索ごとに単語間遷移に相当する仮説を展開する必要がある．単語に比べて構成単位の小さいサブワードではサブワード連鎖の探索を頻繁に繰り返すことになり，デコード処理に大きな負荷がかかる．また，多くの研究では語彙外の単語をすべて 1 つの未登録語クラスとして扱っているため，特定のカテゴリに属する未登録語の持つ単語間の制約や音韻系列の言語的特徴などを十分にモデル化するまでには至っていない．

† ATR 音声言語コミュニケーション研究所
ATR Spoken Language Translation Research Labs.

我々は未登録語を含む音声の高精度な認識を可能にすることを目的として，クラス依存サブワードを用いた言語モデルを提案している[6),7)]．本言語モデルは，単語クラス N-gram[8)]と未登録語に対応する複数のサブワードモデルから構成される．これらのサブワードモデルは未登録語の語彙クラス（たとえば，人名クラス，地名クラスなど）に依存して構築され，単語クラス N-gram の下位階層としてモデル化される．このように言語モデルを階層化することによって，クラス N-gram による単語連鎖の制約と，未登録語を特徴付ける音韻並びの制約とをお互いの干渉なく統合することが可能となる．また，サブワードモデルは未登録語の語彙クラスごとに対象を限定することで，読みの統計的特徴をより高精度にモデル化することが期待できる．

これまで，クラス依存サブワードモデルは個々のサブワードを単語と見なした N-gram 形式で実装していた[6)]．このような実装はクラス N-gram を扱える従来のデコーダであれば，修正なしに利用可能であるというメリットがある．しかしながら，階層的言語制約によって構築されているモデルに対して，階層化を意識しないデコーダを用いることは，単語連鎖とその下位階層であるサブワード連鎖とをフラットなモデルとして扱うことを意味する．このため，言語モデルサイズの肥大化やデコードのオーバヘッドにともなう計算負荷の増大という副作用があった．この副作用を解消するためにも，モデルの特長を生かした新たなデコーダが望まれる．

今後，地名，施設名など，人名以外の未登録語クラスに関しても対象を拡張していくためには，モデルのコンパクト化やデコーダによる仮説探索処理の効率化とともに，それぞれの未登録クラスのより高精度なモデル化が必要になっていく．日本人の姓/名を未登録語の対象としたサブワードモデルのモデル化には，これらの持つ言語的知見に基づいて，単一モーラとモーラ連鎖をサブワード単位として定義し，モーラ並び，およびモーラ長を特徴量としてきた[6)]．ところで，日本人姓/名の言語的な傾向として特徴的と考えられるものは，モーラ長以外にもあげることはできる．サブワードモデルをより高精度にモデル化するためにも，特定のクラスに属する未登録語に対して特徴的な傾向を持つパラメータについて調査しておくことは検討に値する．

本論文では，2 章で日本人姓/名を対象としたサブワードモデルのモデル化について述べる．3 章では，サブワードモデルを用いた階層化言語モデルをより効率的に探索することを目的とした，新しいデコーダの実装法を提案する．本デコーダでは，単語連鎖の下位階層であるサブワード連鎖をデコードするためにサブワードネットワークを採用した．このことにより仮説のデコードを階層化し，効率的な処理を可能としている．また，提案したデコーダの構造を応用し，日本人姓/名を対象としたモデル化に有効と思われる言語的特徴量について検討した．4 章では評価実験を行い，提案したデコーダの性能を評価した．また，デコーダの構造に基づいて作成したサブワードモデルに対して，パープレキシティに基づいた指標で比較するとともに，未登録語を含む音声認識実験を行い，提案手法の有効性を検証した．

2. 人名を対象としたサブワード モデル

2.1 クラス依存サブワード モデル

単語クラス N-gram[8)]では，単語系列 $W = \{w_1, w_2, \ldots, w_N\}$ の言語尤度 $p(W)$ は次式で表すことができる．

$$p(W) = \prod_{i=1}^{N} p(w_i|C^{w_i}) \cdot p(C^{w_i}|C^{w_{i-1}}) \qquad (1)$$

ただし，w_i は i 番目の単語，C^{w_i} は単語 w_i の語彙クラスを表す．ここで，単語 w が未登録語である場合，数式 (1) 第 1 項のクラス内単語生起確率は次式となる．

$$p(w|C^w) = p(M^w|C^{oov}) \qquad (2)$$

ただし，C^{oov} は単語 w の属する未登録語クラス，$M^w = \{m_1, m_2, ..m_L\}$ は単語 w のモーラ系列である．

図 1 にクラス依存サブワードモデルに基づく言語モデルの概念図を示す．上層のクラス N-gram では，数式 (1) の $p(C^{w_i}|C^{w_{i-1}})$ で定義される単語間の制約をかけ，下位階層のサブワードモデルでは，数式 (2) で定義される未登録語クラス内でのモーラ並びの制約をかける．また，サブワードモデルの制約は下位の階層に隠蔽されるため，上層のクラス N-gram とは干渉し

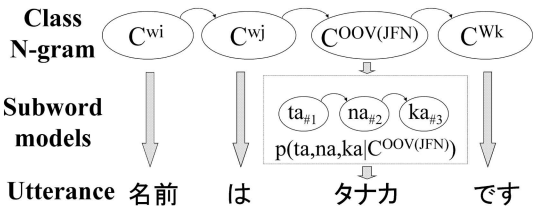

図 1 未登録語認識のための言語モデル
Fig. 1 Language models for OOV word recognition.

ない．

2.2 モーラ長を特徴量に加えたモデル化

日本人姓/名を対象にしたサブワードモデルのより精緻なモデル化を目的として，日本人姓/名に特有の言語的特徴を表現する特徴量について考える．これまでに日本人姓/名の言語的特徴について，モーラの連鎖とモーラ長に特有の傾向があることが分かっており，サブワードモデルはこれらの知見に基づき，サブワード bigram にモーラ長の特徴量を加えてモデル化されていた[6]．

ここで，モーラ長を特徴量に加えたサブワードモデルを定式化する．単一モーラおよびモーラ連鎖をサブワード $s_i = \{m_{l_{i-1}+1}, .., m_{l_i}\}$ として定義すると，

$$\begin{aligned} M^w &= \{m_1, m_2, \ldots, m_L\} \\ &= \{(m_1, .., m_{l_1}), (m_{l_1+1}, .., m_{l_2}) \\ &\quad , \ldots, (m_{l_{K-1}+1}, .., m_{l_K})\} \\ &= \{s_1, s_2, \ldots, s_K\} \end{aligned} \tag{3}$$

となる．ここで l_i は i 番目のサブワード s_i の最終モーラ m_{l_i} の位置を示す（$1 \le l_1 < \ldots < l_{i-1} < l_i < l_{i+1} < \ldots < l_K = L$）．数式 (2) を未知語クラス C^{oov} におけるサブワード bigram $p_{oov}(s_i|s_{i-1})$ の積として展開すると，

$$\begin{aligned} p(M^w|C^{oov}) &= p_{oov}(s_1) \\ &\quad \cdot \prod_{i=2}^{K} p_{oov}(s_i|s_{i-1}) \end{aligned} \tag{4}$$

となる．このとき，サブワードの特徴量としてモーラ長を追加した場合を考える．

$$\begin{aligned} p_{oov}(l > l_i) &= p_{oov}(l > l_i | l > l_{i-1}) \\ &\quad \cdot p_{oov}(l > l_{i-1}) \end{aligned} \tag{5}$$

となることを利用し，C^{oov} クラス内の単語のモーラ長が l_K となる確率 $p_{oov}(l = l_K)$ を展開すると，

$K = 1$ の場合

$$p_{oov}(l = l_K) = p_{oov}(l = l_1) \tag{6}$$

$K > 1$ の場合

$$\begin{aligned} &p_{oov}(l = l_K) \\ &\quad = p_{oov}(l > l_1) \\ &\quad \cdot\ p_{oov}(l > l_2 | l > l_1) \\ &\quad \cdot\ p_{oov}(l > l_3 | l > l_2) \\ &\quad \ldots \\ &\quad \cdot\ p_{oov}(l = l_K | l > l_{K-1}) \end{aligned} \tag{7}$$

となる．したがって，数式 (4) にモーラ長の確率を追加した場合，以下の式に展開できる．

$K = 1$ の場合

$$\begin{aligned} &p(M^w \mid C^{oov}) \cdot p_{oov}(l = l_K) \\ &\quad = p_{oov}(s_1) \cdot p_{oov}(l = l_1) \end{aligned} \tag{8}$$

$K > 1$ の場合

$$\begin{aligned} &p(M^w \mid C^{oov}) \cdot p_{oov}(l = l_K) \\ &\quad = p_{oov}(s_1) \cdot p_{oov}(l > l_1) \\ &\quad \cdot \prod_{i=2}^{K} p_{oov}(s_i|s_{i-1}) \cdot p_{oov}^{L}(i) \end{aligned} \tag{9}$$

ただし，

$$\begin{aligned} &p_{oov}^{L}(i) \\ &\quad = \begin{cases} p_{oov}(l > l_i | l > l_{i-1}) & \text{if } i < K \\ p_{oov}(l = l_i | l > l_{i-1}) & \text{if } i = K \end{cases} \end{aligned} \tag{10}$$

数式 (9) に基づいて，サブワード s_{i-1} からサブワード s_i への遷移確率 $p_{oov}(s_{i-1} \to s_i)$ を以下に定義する．

- 遷移先が単語の終端以外 ($i < K$)

$$\begin{aligned} p_{oov}(s_{i-1} \to s_i) &= p_{oov}(s_i|s_{i-1}) \\ &\quad \cdot p_{oov}(l > l_i | l > l_{i-1}) \end{aligned} \tag{11}$$

- 遷移先が単語の終端 ($i = K$)

$$\begin{aligned} p_{oov}(s_{i-1} \to s_i) &= p_{oov}(s_i|s_{i-1}) \\ &\quad \cdot p_{oov}(l = l_i | l > l_{i-1}) \end{aligned} \tag{12}$$

このように，サブワード間の遷移確率は同じサブワード間の遷移であっても，遷移先のモーラ位置，単語の終端か否かの属性により異なる値を持つことになる．このためサブワードモデルの作成に際しては，これらの属性ごとにエントリを展開しモデル化される．

3. サブワードモデルの実装

3.1 単語 N-gram 形式による実装

これまでは，サブワードモデルをクラス N-gram 形式に対応したデコーダで扱えるように，以下のような実装を行っていた[6]．

サブワードの単位として用いるモーラおよびモーラ連鎖は擬似的な単語として扱い，認識辞書およびクラス N-gram に組み込む．その際，各サブワード単位は以下のラベル付けによる展開を行い，ラベル違いの同一サブワード単位を複数生成する．ラベルは，a) 未登録語クラス（日本人姓/名），b) 単語内での開始モーラ位置，c) 単語の終端か否か，の 3 項組である．ラベル付きのサブワードに関するモデル間の遷移は次のような制約を受ける．

- 登録単語クラスからサブワードへの遷移は開始モーラ位置が 1 のラベルを持つ場合のみ．
- サブワードから登録単語クラスへの遷移は終端ラベルのついた場合のみ．
- サブワード間での遷移はモーラ位置が連接し，か

つ同じ未登録語クラスに属する場合のみ．

これらの制約から外れる遷移は N-gram の遷移確率を明示的に 0 にすることで制限をかける．このため，クラス bigram は back-off smoothing などの平滑化により圧縮されているデータをクラス数 × クラス数のマトリックスに展開して保持しておくことが要求される．さらに，ラベルの異なるサブワード単位をそれぞれ単独のクラスに割り当てていることにより，クラス数の増加が見込まれる．いま，単語クラスの先行クラス（from class）数を $N_{C_w^f}$，後続クラス（to class）数を $N_{C_w^t}$，単一モーラの種類を N_m（日本人姓/名の合計），モーラ長の最大長を L_{max} とすると，単一モーラの先行クラス数は

$$N_{C_m^f} = N_m \cdot L_{max} \tag{13}$$

後続クラス（to class）数は，単語終端と非終端で 2 種類必要なため，

$$N_{C_m^t} = 2 \cdot N_m \cdot L_{max} \tag{14}$$

となる．したがって，サブワードモデルの追加によるクラス数は，

$$N_{C_{w+sw}^f} = N_{C_w^f} + N_m \cdot L_{max} \tag{15}$$

$$N_{C_{w+sw}^t} = N_{C_w^t} + 2 \cdot N_m \cdot L_{max} \tag{16}$$

となり，クラス bigram サイズの比は次式で計算できる．

$$\frac{N_{C_{w+sw}^f} \cdot N_{C_{w+sw}^t}}{N_{C_w^f} \cdot N_{C_w^t}} = \frac{(N_{C_w^f} + N_m \cdot L_{max}) \cdot (N_{C_w^t} + 2 \cdot N_m \cdot L_{max})}{N_{C_w^f} \cdot N_{C_w^t}} \tag{17}$$

たとえば，作成するモデルの緒元を $N_{C_w^f} = N_{C_w^t} = 700$，$N_m = 190$，$L_{max} = 9$ とする．数式 (17) に基づき計算すると，サブワードモデルを追加することにより bigram のサイズは約 20 倍に肥大化する．実際にはモーラ二連鎖をサブワードエントリに追加するため，言語モデルのサイズはさらに大きくなると予想される．

また，音声認識時のデコードに関しても，効率の低下が生じる．サブワード間のデコードでは，モデルの持つ終端ラベルやモーラ位置の制約により遷移の許されるノードは自ずと限定される．しかるに，サブワードを単語として登録している場合には，デコーダはサブワードと登録単語とを区別しない．**図 2** に単語 N-gram 形式で実装した場合の lexical tree の模式図を示す．この図のように，モーラ位置などの属性でラベル付けされたサブワードの各エントリは登録単語とともに lexical tree 上に展開される．デコード処理では，サブワード単位ごとに lexical tree の先頭ノードから終端ノードへとノード間遷移を繰り返しながら仮説が展開されていく．tree の終端ノードに到達した時点で 1 つのサブワード仮説が生成され，tree の先頭ノードに遷移することで後続するサブワードのデコードに移行する．このとき，すべての単語に接続の可能性が許されているため，tree 内での仮説展開に比べて探索空間が大きく拡大する．サブワードは通常の単語に比べて短い単位から構成されているため，この負荷の大きいサブワード間の遷移を頻繁に繰り返すことになる．本来，サブワード間の制約を受けて制限された空間のみを探索すればよいにもかかわらず，単語間遷移にみられる広い仮説空間を探索することは非常に効率が悪く，処理量増加の原因となっている．

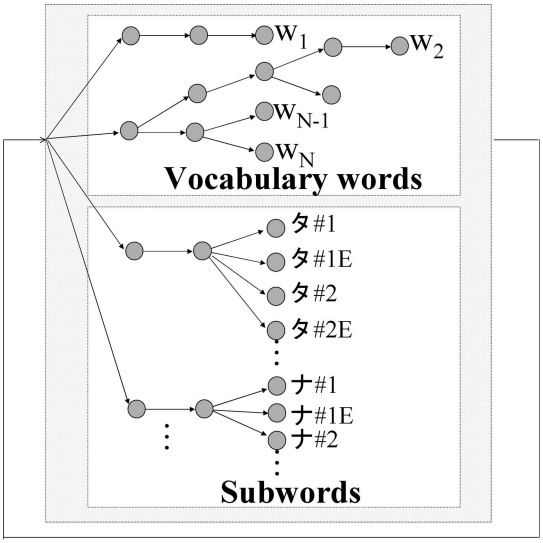

図 2 N-gram 形式で実装した場合の lexical tree
Fig. 2 Lexical tree for N-gram implementation.

3.2 モーラ長制約を省略した実装

単語 N-gram ベースで構築したサブワードモデルの肥大化の原因は，各サブワードエントリをモーラ位置などのラベルごとに展開してモデル化したことにある．デコーダの修正を行うことなくシステムのコンパクト化を図るためには，サブワードエントリからラベルごとの展開をなくせばよい．そこで，数式 (11)，(12) で定義していたサブワード間の遷移確率からモーラ長確率に関する項を外して遷移確率を定義する．

$$p_{oov}(s_{i-1} \rightarrow s_i) = p_{oov}(s_i | s_{i-1}) \tag{18}$$

このように実装した場合の lexical tree の模式図を **図 3** に示す．図 2 と比較してサブワードモデルの終端ノードが減少し，デコーダの探索効率の改善が期待

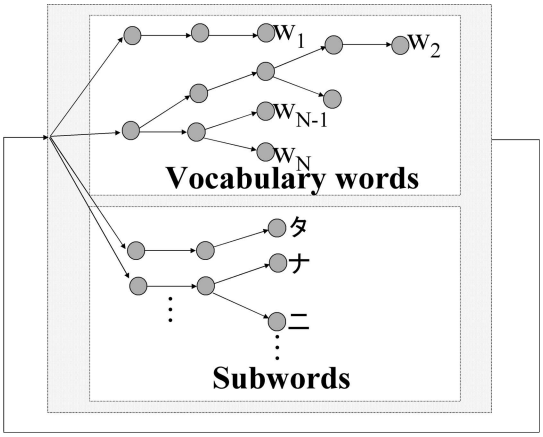

図 3 N-gram 形式で実装した場合の lexical tree（モーラ長制約を省略）
Fig. 3 Lexical tree for N-gram implementation (Without mora durational condition).

できる．また，言語モデルの bigram のサイズに関しては，

$$N_{C^f_{w+sw}} = N_{C^f_w} + N_m \quad (19)$$

$$N_{C^t_{w+sw}} = N_{C^t_w} + N_m \quad (20)$$

として前節と同様に計算すると，サブワードモデル追加による bigram のサイズの増加は約 2 倍とモーラ長を採用したモデルに比べて 1/10 のサイズとなる．ただし，このような実装はサブワードモデルの簡略化にともなう認識性能の劣化が懸念される．この認識性能に関する検討については 4 章で評価実験を行う．

3.3 デコードの階層化によるサブワード遷移確率の分離

サブワードモデルからモーラ長制約を省略することなく言語モデルのサイズ削減とデコードの効率化を実現するために，デコーダの見直しを行った．

3.1 節で示したように，従来の実装ではサブワードの遷移制約から外れた遷移に対して明示的に遷移確率 0 を与える必要があった．このことにともなう言語モデルの肥大化を避けるために，デコードを階層化し，クラス N-gram 言語モデルからサブワードの遷移確率を分離する．また数式 (9) で明らかなように，モーラ長にかかわる確率はモーラの種類によらないため，サブワード bigram とモーラ長に関する確率とは独立にモデル化する．

サブワードモデルを単語 N-gram 形式で実装する場合，個々のサブワードを単語と見なしてモデル化していた．実際には図 1 に示したように，単語連鎖の下位階層としてサブワードモデルが存在しており，サブワード系列の仮説を生成する過程においては上層のクラス N-gram への遷移を考慮する必要はない．そこで，単語仮説を生成する lexical tree とは独立に，サブワード系列の仮説を探索するためのネットワークを作成する．

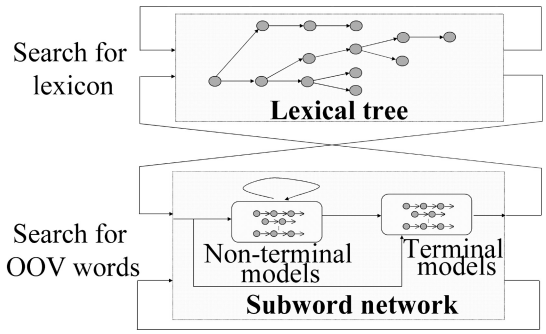

図 4 サブワードネットワークを用いたデコード
Fig. 4 Decoding modules using subword network.

図 4 にサブワードネットワークを用いたデコードの模式図を示す．サブワードネットワークでは，個々のサブワードに対応するモデルを作成するにあたり，単語末に位置する終端モデルとそれ以外に位置する非終端モデルの 2 種類を用意する．非終端モデルからは非終端モデルと終端モデルのどちらか一方への遷移が可能である．終端モデルに到達した時点でサブワード系列のデコードが完了する．lexical tree のレイアとサブワードネットワークとのレイアを分離して実装することにより，レイア間の遷移はお互いの終端ノードからの遷移に限定される．このため，サブワード系列の探索空間と登録単語の探索空間とはお互いに干渉することはなく，無駄な仮説の展開が抑制され効率的なデコードが可能となる．

3.4 デコーダの構造を応用した特徴量の導入

これまで，日本人姓/名を対象としたサブワードモデルでは，サブワード bigram とモーラ長を特徴量に用いてきた．ところで，モーラ長以外にも日本人姓/名の特徴的な言語的傾向をあげることは可能である．たとえば，“ザ–ワ”，“グ–チ” など単語の終端に集中して存在するサブワードがある一方で，“ア–オ”，“キ–ク” などは単語の終端にはほとんど現れない．このような単語終端位置での生起確率の片寄りも，日本人姓/名を対象としたサブワードモデルのモデル化に有効な特徴量と思われる．

実装したサブワードネットワークは，単語の終端に位置する終端モデルとそれ以外に位置する非終端モデルの 2 種類を個別に配置した構造を持っている．この構造を利用することで，単語終端位置での生起確率を特徴量に追加したモデルは容易に定式化できる．サブ

表 1 辞書および言語モデルの比較
Table 1 Comparison of lexicon and language models.

	N-gram 実装	提案法
Lexicon（約 18,000 語）	2.4 MB	1.0 MB
Class N-gram（700 class）	150.9 MB	2.2 MB
Subword models（x2）	-	0.5 MB
Total	153.3 MB	3.8 MB

ワード s_i がクラス C^{oov} に属する単語中の語末に存在する確率 $p_{oov,E}(s_i)$ をサブワードの終端確率として定義すると，サブワード s_{i-1} からサブワード s_i への遷移確率は以下のようになる．

- 非終端モデルへの遷移確率

$$p_{oov}(s_{i-1} \to s_i) = p_{oov}(s_i|s_{i-1}) \cdot (1 - p_{oov,E}(s_i)) \quad (21)$$

- 終端モデルへの遷移確率

$$p_{oov}(s_{i-1} \to s_i) = p_{oov}(s_i|s_{i-1}) \cdot p_{oov,E}(s_i) \quad (22)$$

したがって，図 4 のサブワードネットワークを用いたデコードにおいて，非終端モデルから非終端モデルへの遷移の際には数式 (21) の遷移確率を用い，非終端モデルから終端モデルへの遷移の際には数式 (22) の遷移確率を用いて尤度計算を行う．

4. 評価実験

4.1 デコーダの比較

単語数約 18,000 語，クラス数約 700 の言語モデルを基本モデルとし，日本人姓/名を未登録語と見なしてサブワードモデルを作成した．サブワードモデルは単一モーラ 95，二連鎖モーラ 150 の計 245 個をサブワードの単位として選択し，サブワード bigram とモーラ長の特徴量に基づいてモデル化した．

言語モデルのサイズについて，従来実装と提案方式との比較を**表 1** に示す．3.1 節で説明したように，従来実装では，a) 未登録語クラス（日本人姓/名），b) 単語内での開始モーラ位置，c) 単語終端か否か，の 3 項組のラベル付けにより展開されたサブワードを擬似的な単語と見なし，個々のエントリに対してそれぞれ個別のクラスを割り当ててクラス N-gram を作成している．このため，クラス N-gram のサイズが肥大化している．一方，サブワードネットワークを用いた提案方式では，サブワードモデルと登録単語との干渉が起こらないため，それぞれのモデルを別々に持つことが可能となる．基本モデルに加えて日本人姓/名の 2 種類のサブワードモデルを追加したサイズの合計は 3.8 MB となり，従来実装に対して言語モデルを 1/40 のサイズに削減できた．

表 2 実験条件
Table 2 Experimental condition.

評価音声	· 旅行会話ドメインの 42 片側会話音声[11]
特徴量	· 16 kHz サンプリング（16 bit） · フレーム周期 10 ms，フレーム長 20 ms · 12 次 MFCC と対数パワー，および それらの一次回帰係数（計 26 次元）
音響モデル	· 音素環境依存 HMnet · 1400 状態 5 混合（男性用モデル） · 1400 状態 15 混合（女性用モデル）
言語モデル	· 単語クラス N-gram[8]+サブワードモデル （詳細は表 1 参照）
デコード	·1 パス 時間同期ビタビサーチ ·2 パス 言語重みを変更したフルサーチ

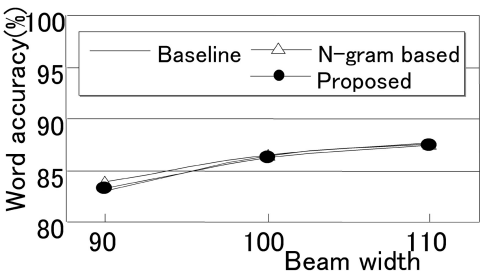

図 5 word accuracy の比較
Fig. 5 Comparison of word accuracy.

次に，未登録語の対象である日本人姓/名を含む評価音声を用いた認識実験を行った．実験条件を**表 2** に示す．評価用音声データには，旅行会話ドメインの 42 片側会話（4,990 単語）を採用した[11]．評価音声に出現する未登録語は，日本人姓 50 語，日本人名 20 語の計 70 語である．実験ではサブワードモデルを含まない基本モデルの結果をベースラインとし，N-gram 形式の従来実装のデコーダ（N-gram based）とサブワードネットワークを用いた提案方式（proposed）との 2 つのデコーダについて，word accuracy と処理量（RTF=real time factor）とで比較した．

図 5 にビーム幅をパラメータとした場合の word accuracy を示す．ベースラインの結果とサブワードモデルを追加した結果とを比較すると，ほぼ同じ word accuracy が得られた．このことはサブワードモデルの導入が未登録語以外の単語認識性能に悪影響を与えないことを示している．また，デコーダによる差もほとんどないことが確認できた．

次に，処理量（RTF）について比較する．RTF の計測に使用した計算機は 2 GB のメモリを搭載した Pentium III（1 GHz）マシンであり，OS は Linux である．RTF は認識処理に費やされた CPU 時間を発声時間で正規化した値であり，実験では全評価用音声に

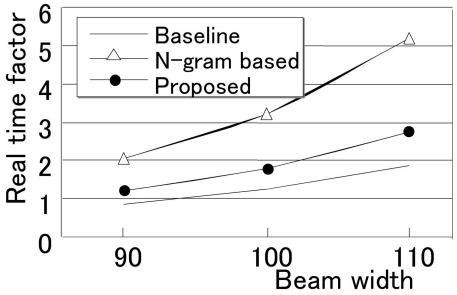

図 6 処理量の比較
Fig. 6 Comparison of processing time.

対する平均値を求めた．実験結果を**図 6** に示す．ビーム幅 110 の条件で比較すると，従来方式は RTF=5.2 であるのに対して，提案方式では RTF=2.8 となり，46%の処理量削減が達成された．

4.2 サブワード モデルの性能評価

3 章では，サブワード モデルの実装に関して，サブワード モデルからモーラ長を省略することで言語モデルをコンパクト化する試みについて述べた．このような実装は，サブワード モデルを簡略化したことによる認識性能の劣化が懸念される．ここでは，モーラ bigram のみの特徴量でモデル化した簡略化サブワード モデルとモーラ長の属性を含むサブワード モデルとの比較実験を行った．また，サブワードの終端確率を特徴量に追加したモデルについでも同時に比較した．サブワード モデルの学習に関しては，約 30 万人の著名人を収録した人名リスト[10]をコーパスとして用いた．このコーパスに基づいて，日本人姓/名に高頻度で出現するモーラ二連鎖を抽出し，単一モーラ（95 種類）と抽出したモーラ二連鎖とをサブワード単位として選択した．このサブワードに対する bigram および，モーラ長，サブワードの終端確率をそれぞれ学習し，特徴量の異なる 3 種類のサブワード モデルを作成した．

(1) サブワード bigram のみでモデル化
(2) サブワード bigram にモーラ長の特徴量を追加したモデル化
(3) サブワード bigram にサブワードの終端確率の特徴量を追加したモデル化

4.2.1 パープレキシティによる比較

作成した 3 種類のサブワード モデルに対してパープレキシティを求めた．評価用の単語セットとして，ATR 音声認識システム SPREC[11]で使用している単語辞書の語彙から日本人姓/名を除いた約 16,000 語をクラス外の評価セット，単語辞書に含まれる日本人姓（約 600 語），日本人名（約 300 語）の語彙をそれぞれクラス内の評価セットとして用いた．

言語モデルの評価指標としてはパープレキシティを用いるのが一般的である．サブワード モデルのモデル化に際しては，未登録語として設定した日本人姓，日本人名に対しては推定精度の良いモデルが望まれるのはもちろんのこと，対象クラス外の語彙に対しては推定精度の悪いモデルを作成することで，未登録クラスに対する識別能力の高いモデル化を実現できる．いい換えれば，目的とするサブワード モデルは，日本人姓/名を評価セットに用いた場合には低いパープレキシティを示し，日本人姓/名以外の語彙を用いた評価セットに対しては高いパープレキシティを示すモデルが望まれる．日本人姓を対象としたサブワード モデルの評価において，日本人姓を評価セットに用いて求めたパープレキシティを PP_{JFN}，日本人姓/名以外の語彙を評価セットに用いて求めたパープレキシティを $PP_{w/oJN}$ とすると，PP_{JFN} は低い値を，$PP_{w/oJN}$ は高い値を示すほど日本人姓と他の語彙との間の分離度の高いことになる．したがって，その比 $PP_{JFN}/PP_{w/oJN}$ の値が低いほど，サブワード モデルとして良いモデル化といえる．そこで今回は，このパープレキシティの比 $PP_{JFN}/PP_{w/oJN}$（対数をとるとエントロピーの差となり，尤度差を示す）をサブワード モデルの評価尺度として用いた．

実験では，サブワード単位として選択するモーラ二連鎖の数を変えながら日本人姓/名それぞれのモデルを作成し，これらのモデルに対してクラス外評価セットで求めたパープレキシティとクラス内評価セットで求めたパープレキシティとの比を計算した．

実験結果を**図 7** に示す．図中 (a) は日本人姓を対象として学習したサブワード モデルに対して，日本人姓の評価セットで求めたパープレキシティと姓/名を外した評価セットで求めたパープレキシティの比を，二連鎖モーラのエントリ数を横軸にとってプロットしたグラフである．(b) は日本人名を対象にした同様のグラフである．日本人姓/名のどちらを対象とした条件についても，サブワード bigram のみでモデル化した場合よりもモーラ長の特徴量を追加したモデル（bigram+duration）の値は小さくなっており，追加した特徴量の効果が確認できる．この結果から，コンパクト化を目的としたサブワード モデルの簡略化は，未登録語に対する認識性能の劣化をともなうことが予想される．

次に，モーラ長を特徴量に追加したサブワード モデルと終端確率を特徴量に追加したモデル（bigram+terminal prob.）とを比較する．二連鎖モーラのエントリ数が小さい条件で，サブワードの終端確

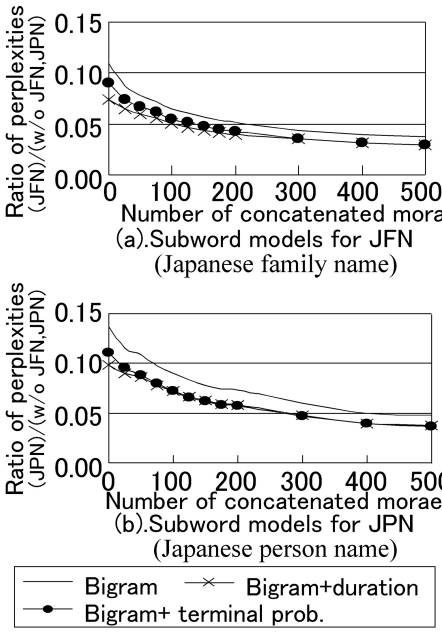

図 7 サブワードモデルの性能比較
（サブワードモデルに対するパープレキシティの比）
Fig. 7 Comparison between subword models (Ratio of perplexities for subword models).

率を特徴量に追加したモデルの値がモーラ長を特徴量に追加した場合よりも大きくなっているのは，終端位置における生起確率の偏りは，単一モーラよりも二連鎖モーラによってより強く特徴付けられているためと考えられる．二連鎖モーラのエントリ数が 200 以上の条件では，モーラ長の特徴量を追加したモデル，サブワードの終端確率の特徴量を追加したモデルともにほぼ同じ値を示していることから，両者のモデルは同等の性能を持つものと期待できる．

4.2.2 未登録語に対する認識性能による比較

サブワードモデルの性能を未登録語に対する音声認識結果で比較する．サブワードモデルは，95 種類の単一モーラに追加する二連鎖モーラのエントリ数を 150，500 の 2 通りの条件でサブワード単位を定義し，それぞれのモデルを作成した．サブワードモデル以外の実験条件は，デコーダの比較実験と同一である．デコーダはサブワードネットワークを用いた提案方式を採用した．

旅行会話ドメインの評価音声データに存在する未登録単語 70 語に対する正解数を**表 3** に示す．ちなみに，日本人姓/名を単語登録した場合の正解数は 47 であった．パープレキシティの比を指標にした図 7 の結果からも予測されるように，二連鎖モーラのエントリ数を 150 から 500 に増やすことで未登録語に対する認識性能は向上する．また，二連鎖モーラのエントリ数を一致させた条件で比較した場合，サブワード bigram のみを特徴量として用いたモデルの正解数に対して，モーラ長を特徴量に追加したモデル，サブワードの終端確率を特徴量に追加したモデル，ともに正解単語数の向上が見られる．二連鎖モーラのエントリ数を 500 にした条件では，サブワード bigram のみでモデル化したときの正解数が 39 であるのに対し，特徴量を追加した 2 種類のモデルを用いた場合は，それぞれ 45，46 と約 15%正解数が増加し，追加した特徴量の有効性が確認できた．この結果は，3.2 節で述べたサブワードモデルの簡略化が未登録語の認識性能の低下をともなうことを実証している．

表 3 未登録語に対する認識結果（70 語中の正解数）
Table 3 Recognition result for OOV words (Number of correctly recognized words in 70 words).

		Number of concatenated morae	
		150	500
Models	bigram	38	39
	bigram+duration	42	45
	bigram+terminal prob.	42	46

5. むすび

本論文では，未登録語を含む音声の認識を可能とするクラス依存サブワードモデルを効率的にデコードするための実装手法について検討した．モーラ長制約を省略したサブワードモデルの簡略化については，未登録語に対する認識性能の劣化をともなうことが分かった．サブワードモデルの特徴量を損なわずに効率的にデコードする実装方法として，サブワードネットワークを用いたデコーダについて提案した．提案方式を従来の単語 N-gram ベースのデコーダと比較した結果，認識性能を劣化させることなく，言語モデルのデータサイズを 1/40 にし，46%の処理量削減が可能であることが明らかとなった．

また，提案したサブワードネットワークの構造を応用し，日本人姓/名を対象としたモデル化に有効と思われる言語的特徴量について検討した．パープレキシティの比を評価基準としたシミュレーション，ならびに評価用音声データに出現する未登録語の認識実験により，異なる特徴量を用いて学習した 3 種類のサブワードモデルを比較した．サブワード bigram のみを用いてモデル化した場合の未登録語の正解数 39 に対して，特徴量としてモーラ長，単語の終端位置でのサブワード生起確率を追加して学習したサブワードモデ

ルでの正解数は，それぞれ 45，46 と約 15%の正解数向上が見られた．この結果，サブワード bigram に対して追加されたこれらの特徴量が効果的であることが実証された．

以上のことから，本論文で提案したデコーダがサブワードモデルの効率的なデコードを実現し，かつサブワード生起確率などの特徴量を追加した高精度なサブワードモデルの導入に対しても容易に実装できることが明らかとなった．

参 考 文 献

1) 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).
2) 甲斐充彦，中川聖一：冗長語・言い直し等を含む発話のための未知語処理を用いた音声認識システムの比較評価，信学論 D-II，Vol.J80, No.10, pp.2615–2625 (1997).
3) Klakow, D., Rose, G. and Aubert, X.: OOV-detection in large vocabulary system using automatically defined word-fragments as fillers, *Proc. Eurospeech1999*, pp.49–52 (1999).
4) Kneissler, J. and Klakow, D.: Speech recognition for huge vocabularies by using optimized sub-word units, *Proc. Eurospeech2001*, pp.69–72 (2001).
5) Bazzi, I. and Glass, J.: Learning units for domain-independent out-of-vocabulary word modeling, *Proc. Eurospeech2001*, pp.65–68 (2001).
6) 谷垣宏一，山本博史，匂坂芳典：未登録語のクラス依存サブワードモデルを用いた音声認識，信学技報，SP99-123, pp.49–54 (1999).
7) Onishi, S., Yamamoto, H. and Sagisaka, Y.: Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes, *Proc. Eurospeech2001*, pp.693–696 (2001).
8) 山本博史，匂坂芳典：接続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, No.11, pp.2146–2151 (2000).
9) 清水　徹，山本博史，政瀧浩和，松永昭一，匂坂芳典：大語彙連続音声認識のための単語仮説数削減，信学論 D-II，Vol.J79, No.12, pp.2117–2124 (1996).
10) 日外アソシエーツ：30 万人読み方書き方辞典，ISBN4-8169-7020-7 (1993).
11) 内藤正樹，山本博史，中嶋秀治，中村　篤，匂坂芳典：対話音声を対象とした連続音声認識システムの試作と評価，信学論 D-II，Vol.J84, No.1, pp.31–40 (2001).

(平成 13 年 11 月 16 日受付)
(平成 14 年 4 月 16 日採録)

小窪　浩明（正会員）

昭和 63 年上智大学理工学部電気電子工学科卒業．平成 2 年同大学院博士前期課程修了．同年（株）日立製作所中央研究所に入所．平成 7〜9 年 ATR 音声翻訳通信研究所研究員．平成 12 年より ATR 音声言語通信研究所に出向．音声認識の研究開発に従事．日本音響学会会員．

大西　茂彦

平成 3 年横浜国立大学大学院修士課程修了．同年日本電信電話（株）LSI 研究所入所．アナログ・ディジタル混載 LSI，音声信号処理モジュールの開発等に従事．現在，ATR 音声言語コミュニケーション研究所研究員．電子情報通信学会，日本音響学会各会員．

山本　博史

昭和 54 年東京大学農学部農業生物学科卒業．昭和 56 年同大学院修士課程修了．同年（株）CSK に入社．平成 8 年より ATR 音声翻訳通信研究所に出向．音声認識の研究開発に従事．電子情報通信学会，言語処理学会，日本音響学会各会員．

菊井玄一郎（正会員）

1986 年京都大学大学院電気工学第二専攻修士課程修了．同年 NTT に入社，2001 年 4 月より（株）国際電気通信基礎技術研究所（ATR）に出向，現在に至る．自然言語処理，音声言語処理，特に自動翻訳，WEB 情報検索，多言語情報検索等の研究開発に従事．ACL，人工知能学会，言語処理学会に所属．