社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (39)
2006／12／22

可変長サブワードHMMに基づく未知語処理を導入した音声認識

本間 真一　小林 彰夫　尾上 和穂　佐藤 庄衛　今井 亨　都木 徹

NHK 放送技術研究所

〒157-8510 東京都世田谷区砧 1-10-11

E-mail: {homma.s-fc, kobayashi.s-fs, onoe.k-ec, satou.s-gu, imai.t-mq, takagi.t-fo}@nhk.or.jp

あらまし 一般に，大語彙連続音声認識では，認識装置に登録可能な語彙のサイズに制限があるため，語彙に登録されていない単語(未知語)は認識できないという問題がある．そこで本稿では，任意の長さのサブワードの連接からなるカナ文字列によってあらゆるパターンの未知語を表現し，認識結果を出力する手法を提案する．未知語を構成する可変長サブワード系列の出力確率は，一般的な長さ1のシンボルを出力するHMMを拡張し，最尤推定によって学習する．また，登録するサブワードの種類を削減するために，MDL規準によるサブワードの選択と出力確率の再推定を行う．認識時には，登録語彙によって構築した言語モデルと未知語用HMMを組み合わせ，未知語を含む音声区間をカナ文字列で出力する．自然ドキュメンタリー番組の音声認識実験の結果，未知語を含む発話の単語誤り率は26.7%から18.4%に改善した．

キーワード 大語彙連続音声認識，言語モデル，未知語処理，HMM，MDL規準

Speech Recognition with Out-of-Vocabulary Word Processing Using a Variable-Length Sub-Word HMM

Shinichi HOMMA　Akio KOBAYASHI　Kazuo ONOE　Shoei SATO　Toru IMAI and Tohru TAKAGI

NHK (Japan Broadcasting Corporation) Science and Technical Research Laboratories

1-10-11 Kinuta, Setagaya-ku, Tokyo, 157-8510 Japan

E-mail: {homma.s-fc, kobayashi.s-fs, onoe.k-ec, satou.s-gu, imai.t-mq, takagi.t-fo}@nhk.or.jp

Abstract General LVCSR has a problem that Out-Of-Vocabulary (OOV) words cannot be recognized because of limitations of registered words. In this paper, we propose a novel approach to recognize every OOV word by using Kana character strings of connected variable-length sub-words. We estimate output probabilities of the sub-word patterns by maximum likelihood estimation applying a general HMM which emits a unit symbol at a time. In order to reduce the number of the sub-words, we select the sub-words based on the MDL criterion and re-estimate their output probabilities. When we perform speech recognition, the HMM for OOV words is used with a language model constructed by using vocabulary words and outputs Kana character strings from the input speech segments including OOV words. In a recognition experiment of a broadcast documentary program dealing with nature, the word error rate of evaluation data including OOV words in each sentence was reduced from 26.7% to 18.4%.

Keyword LVCSR, Language Model, OOV, HMM, MDL Criterion

1. はじめに

NHKは，音声認識技術を利用した自動字幕制作システムを開発し，ニュース，音楽，スポーツ等の生放送番組において，リアルタイムの字幕放送を実現した[1][2]．また，将来のサーバ型放送等で用いるメタデータの生成のため，音声認識を活用する研究も行っている[3]．こうした音声認識の研究や実用化の進展に伴い，認識対象がニュース以外の一般番組にも拡大してきており，認識対象の語彙を拡大する必要が生じてきた．ニュース番組では類似した記者の原稿を利用できることや，スポーツ番組では話題が限定できることなどから，語彙は2～4万単語程度としていたが，様々な話題を扱う情報番組などでは，語彙を10万単語規模に拡大することが有効である[4]．しかし，実世界の語彙サイ

ズには制限がないため，いくら認識対象の語彙サイズを大きくしても，エントリに存在しない単語(未知語)は認識できないという問題は解消できない．

従来の未知語への対処を考慮した音声認識の手法には，未知語を任意の音素や音節の系列とみなして，音素 N-gram[5]や音節 HMM[6]を用いて認識する方式があるが，言語モデルによる単語内や単語間の制約が効きにくいため，得られる認識精度に限界がある．一方，カナで表記した単語の読みを学習テキストに加えて構築した言語モデルによって未知語をカナ文字列とみなして認識する方式[7][8]や，未知語を単語より短いサブワードの単位に分割したテキストで学習した言語モデルによって認識する方式[9]が提案されたが，学習テキストに存在しないパターンの未知語は認識できないという問題がある．また，学習テキストから選択された高頻度の音節列2-gramを用いる方式[10]があるが，あらゆるパターンの音節列 2-gram を学習することは現実的に困難であることから，この場合も，学習パターンに当てはまらない未知語は認識できないという問題が残る．

放送音声の認識では，時を追うごとに新しい単語が生じること，未知語の前後にも認識誤りが波及しやすいこと，未知語であってもせめてカナで出力してほしいという字幕放送の送出現場の要望があること，カタカナ語の未知語の割合が大きいこと[4]などから，過去に生成された学習テキストから，あらゆる未知語のパターンを予測し，カナによって出力可能にしておくことは重要である．また，NHK のリアルタイム字幕制作システム[11]には，カナ漢字変換機能によってオペレータが同音異義語を修正する機能があるため，これを利用して，カナによる認識出力を，漢字交じりの正しい表記に変換することも可能である．そこで本稿では，未知語を可変長のカナ文字列の組み合わせによって表現する確率モデルを作成することにより，あらゆるパターンの未知語を受理可能な認識手法を提案する．この確率モデルは，一般的な隠れマルコフモデル(HMM)を拡張したものであり，前向き・後向きアルゴリズムを利用してパラメータの学習を行う．この学習の結果，膨大な数のパラメータが得られる可能性があるが，最小記述長(MDL: Minimum Description length)規準[12]を用いて最適なパラメータの選択を行うことにより，不要なパラメータを削減する．また，認識の際に使用する言語モデルは，単語 N-gram モデルであるが，未知語を認識するために，未知語クラス(UNK)の N-gram 確率と本 HMM を組み合わせて利用する．以下では，本手法を具体的に説明するとともに，音声認識実験の結果などから，その効果を考察する．

2. 未知語処理を導入した音声認識

2.1. サブワードによる未知語モデル

従来の音声認識では，認識装置に登録可能な語彙サイズに制限があるため，あらゆる単語を登録することは不可能である．そこで，学習テキスト中で高頻度の単語や，一般によく知られた単語だけを語彙として登録することになるため，特殊な専門用語や固有名詞等は未知語となるケースが多くみられる．たとえば，動物番組の音声認識では，特殊な動物名が未知語となることが多いが，ここでは，例として，「オオアリクイ」「オオスズメバチ」「オオサンショウウオ」「クロサンショウウオ」という４つの動物名が未知語となったことを仮定する．これらの単語を認識可能にするために，「オ」「ア」「リ」「ク」「イ」などの音節の単位を語彙に登録する手法が考えられるが，実際に認識実験をしてみると，これだけではあまり認識精度が改善しない．そこで，音節よりも長く単語よりも短い，単語の一部(サブワード)を語彙に登録することを考える．ここで，サブワードとして採用する文字列は，「よく現れそうな文字列の単位」であり，上記の例では，直感的に「オオ」「アリクイ」「スズメバチ」「サンショウウオ」「クロ」を採用するべきではないかと考えられる．このようなサブワードの単位を，人間が経験的に定めるのではなく，可変長のサブワード系列を出力する HMM(サブワード HMM)を利用して，学習データから統計的に最適なものを取得することにする．なお，音声認識の際に，発音辞書において文字列と音素の対応付けが必要となるため，サブワードの最小単位は１音節を表現するカナ１文字，または，拗音・長音を連結したカナ文字列(例：アー，キャ，キャー)とする．

2.2. サブワード HMM を加えた言語確率

言語モデルとして単語 N-gram モデルを用いる場合，h を単語 w 直前の(N-1)-gram 履歴とすると，言語確率は $P(w|h)$ により求められるが，w が未知語である場合を考慮して，これを次式のように拡張する．

$$P(w|h)=\begin{cases}P_{N-gram}(w|h) & if \quad w\in V_{N-gram}\\ P_{N-gram}(UNK|h)P_{HMM}(w) & else\end{cases} \tag{1}$$

ここで，V_{N-gram} は N-gram 言語モデルの語彙(認識装置の登録語彙)，UNK は未知語クラス，$P_{N-gram}(\cdot)$ は N-gram モデルが単語を出力する確率，$P_{HMM}(\cdot)$ はサブワード HMM がサブワード列を出力する確率を表す．

たとえば，ある未知語 w がシンボル(音節)列 $o_1o_2o_3$ であるとした場合，１つのサブワードを[・]で括って表すと，$[o_1][o_2][o_3]$，$[o_1][o_2o_3]$，$[o_1o_2][o_3]$，$[o_1o_2o_3]$ という４つのサブワード系列を考えることができる．認識時には，次式のように最大の確率値を与えるサブワード系列(最尤サブワード系列)の値を $P_{HMM}(w)$ として採用する．

$$P_{HMM}(w)=\max\begin{Bmatrix}P_{HMM}([o_1])P_{HMM}([o_2])P_{HMM}([o_3])\\ P_{HMM}([o_1])P_{HMM}([o_2o_3])\\ P_{HMM}([o_1o_2])P_{HMM}([o_3])\\ P_{HMM}([o_1o_2o_3])\end{Bmatrix} \tag{2}$$

2.3. 可変長サブワード HMM の学習

ここで考える HMM は，式(2)で示したとおり，単語に含まれる任意のサブワードの系列を出力するものであり，どの系列がどの状態から出力されるかは確率的に決定される．よって，各状態からは，可変長のサブワードが出力されることを想定する．一般的な HMM では，各時刻で１つのシンボルの出力を考えるが，本 HMM では，各時刻で複数の連結したシンボルを出力可能である．以下に，本 HMM の学習方法を具体的に

述べる．

まず，言語モデルの学習用のテキストから収集した単語の読みをカナに変換し，得られたカナ表記の単語集合 V に含まれるすべての単語 w をサブワード HMM の学習データに使用する．

単語 w が，シンボル(音節)列 $o_1^{T_w}=o_1\cdots o_t\cdots o_{T_w}$ (ただし，T_w は w のシンボル長)であるとした場合，ある HMM から w が出力される場合の確率は $P_{HMM}(o_1^{T_w})$ と表される．これを V に含まれるすべての単語について考慮した式は，$\prod_{w\in V}P_{HMM}(o_1^{T_w})$ と書けるが，以下では，この式を最大化する HMM を求めることを考える．なお，ここでは，状態 0～状態 S の $S+1$ 個の状態をもつ left-to-right 型の HMM を想定する．初期状態は 0($t=0$)，最終状態は S($t=T_w$)とし，初期状態と最終状態を除く各状態からは，シンボルとして単語 w のサブワード s が出力されるものとする．なお，単語 w のシンボル長が 1 である場合もあるため，トポロジーにおいて，状態スキップを許容するものとする．

HMM は，前向き・後向きアルゴリズムによって，状態遷移確率とシンボルの出力確率分布を求めることができる．このアルゴリズムによって得られる単語 w の前向き確率 α^w，後向き確率 β^w，時刻 t において状態 i から状態 j の遷移が生じる確率 $\gamma_t^w(i,j)$ の推定式，および，状態 i から状態 j への状態遷移確率 a_{ij} と状態 j におけるサブワード s の出力確率 $P_j(s)$ の再推定式は次のようになる．なお，N はサブワード s のシンボル長(音節数)の最大値を表す．

α の初期化:

$$\alpha_t^w(j)=\begin{cases}1 & if\quad t=0, j=0\\ 0 & if\quad t=0, j\neq 0\end{cases}$$

時刻 $t=1,\cdots T_w$，状態 $j=1,\cdots,$S:

$$\alpha_t^w(j)=\sum_{i=0}^{j}\sum_{\tau=t-N}^{t-1}\alpha_\tau^w(i)P_j(o_{\tau+1}^t)a_{ij} \tag{3}$$

β の初期化:

$$\beta_t^w(i)=\begin{cases}1 & if\quad t=T_w, i=S\\ 0 & if\quad t=T_w, i\neq S\end{cases}$$

時刻 $t=T_w-1,\cdots,0$，状態 $i=0,\cdots,$S:

$$\beta_t^w(i)=\sum_{j=i}^{S}\sum_{\tau=t+1}^{t+N}\beta_\tau^w(j)P_j(o_{t+1}^\tau)a_{ij} \tag{4}$$

$$\gamma_t^w(i,j)=\frac{\sum_{n=1}^{N}\alpha_{t-n}^w(i)P_j(o_{t-n+1}^t)a_{ij}\beta_t^w(j)}{\alpha_{T_w}^w(S)} \tag{5}$$

$$\tilde{a}_{ij}=\frac{\sum_{w\in V}\sum_{t=1}^{T_w}\gamma_t^w(i,j)}{\sum_{w\in V}\sum_{t=1}^{T_w}\sum_{j'=1}^{S}\gamma_t^w(i,j')} \tag{6}$$

$$\tilde{P}_j(s)=\frac{\sum_{w\in V}\sum_{t=1}^{T_w}\sum_{n=1}^{N}\sum_{i=0}^{S}\left[\frac{\delta_s(t,n)\alpha_{t-n}^w(i)P_j(s)a_{ij}\beta_t^w(j)}{\alpha_{T_w}^w(S)}\right]}{\sum_{w\in V}\sum_{t=1}^{T_w}\left[\frac{\alpha_t^w(j)\beta_t^w(j)}{\alpha_{T_w}^w(S)}\right]} \tag{7}$$

ただし，$\delta_s(t,n)$ は以下で定義される．

$$\delta_s(t,n)=\begin{cases}1 & if\quad o_{t-n+1}^t=s\\ 0 & else\end{cases}$$

式(7)の分母は，全学習データ($w\in V$)の全時刻($1\le t\le T_w$)を考慮したときの状態 j に滞在する回数の期待値，分子は，状態 j において，サブワード s を含む出力系列を経て単語 w を出力する回数の期待値を表している．なお，$P_j(s)$ の初期値は，学習データ中の各 s の相対頻度を用いる．

図 1 に，$\alpha_t^w(j)$ を計算するためのトレリスの例($N=2$ の場合)を示す．なお，本図の横軸は時間経過，縦軸はモデルの各状態を示す．一般的な前向きアルゴリズムの計算式では，時刻 t において各状態から出力される記号は，長さ 1 の o_t だけを考慮するのが普通であるが，式(3)では，時刻 t において，各状態から $o_{t-N+1}^t,\cdots,o_t^t$ という長さ N 以下のすべてのサブワードが出力されることを考慮する点に特徴がある．

ところで，本手法と類似したものに N-multigram モデル[13]があるが，この手法は，状態 i において固定長 i のシンボル列が出力され，すべての状態遷移確率は等確率であるエルゴディックモデルを定義する．一方，本手法では，各状態から出力されるシンボル列(サブワード)は，長さ N 以下の可変長の任意のシンボル列であり，状態遷移確率を等確率に限定していない点が異なる．これにより，未知語中のサブワードの位置によって異なる統計的性質を表現することが期待できる．

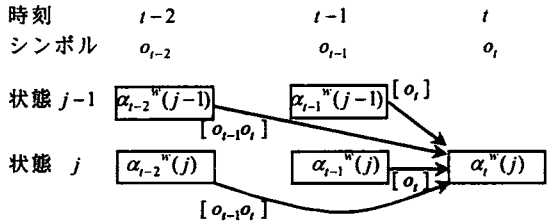

図 1 $\alpha_t(j)$ の計算

2.4. MDL によるパラメータ数の削減

式(7)により，単語 w から生成可能な長さ N 以下のすべてのサブワードの出力確率が求まるが，学習データ

が大量にある場合は，これをそのまま HMM のパラメータに採用すると，パラメータ数が膨大なものになってしまう．そこで，一般的な情報量規準である MDL を利用することにより，HMM のパラメータとして最適なサブワードを選択することを考える．

学習データ w_i の個数を $|V|$，HMM のパラメータ数を m とすると，MDL は以下の式(8)で表され，この式の値が最小となるときに，最適なモデルのパラメータの選択がなされることになる．

$$MDL = -\sum_{i=1}^{|V|} \log P_{HMM}(w_i) + \frac{m}{2} \log |V| \quad (8)$$

なお，ここで考えるパラメータは HMM の出力確率値とする．

以下に，パラメータ数の削減の手順を記す．

1. 学習データから式(6)と式(7)によって，HMM の各パラメータの確率値を推定する．
2. 1 で得られたサブワードの出力確率値のうち，確率の閾値より大きい上位 m 個のサブワードを選択する。
3. 2 で選択された m 個のサブワードの出力確率値の和が 1 になるように，出力確率値の正規化を行う．
4. 3 で得られたサブワードの出力確率を初期値として，式(6)と式(7)により，HMM の各パラメータの確率値を再推定する．
5. 式(8)によって，MDL を計算する．
6. パラメータ数 m を変えながら，1～5 の手順を繰り返して，m の値と MDL の関係を求め、MDL が極小となるときの m の値を見つける．
7. 6 で得られた m 個のサブワードを，MDL から得た最適なモデルのパラメータとみなす．

なお，あらゆる未知語を受理可能とするために，1 音節の全パターンは必ずサブワードとして選択しておくことにする．

3. 実験

3.1. 評価データ

実験に用いるタスクは，自然ドキュメンタリー番組の放送音声とし，評価データには NHK の「地球ふしぎ大自然」の放送音声から，3 日分（2004 年 5 月 10 日，2005 年 7 月 24 日，2005 年 8 月 25 日放送，1180 文 12.3K 単語）を選んだ．この音声の大部分はナレーションであり，複数名の男女によって発話が行われている．この音声には，音声認識を困難にさせる背景雑音や背景音楽が多く混入している．本番組には，クローズドキャプションが付加されているため，字幕用のテキストを得ることができる．しかし，この字幕テキストは，オープンキャプションが存在するシーンにおいて欠落がみられ，かつ，読みやすいように整形されているため，必ずしも忠実に放送音声が字幕化されているものではない．クローズドキャプションが存在しない区間の音声認識結果は，メタデータ(主に動植物名等のキーワード)の自動生成に重要であるため，今回の実験では不完全なクローズドキャプションを言語モデルと音響モデルの学習に利用する．クローズドキャプションの字幕テキストを言語モデルの学習データに追加しただけでは，この区間は未知語となるケースが多いため，提案手法による改善が期待できる．

3.2. 学習テキスト

モデル学習用のテキストには，1995 年 4 月～2005 年 9 月に NHK の「生きもの地球紀行」と「地球ふしぎ大自然」（約 40 分×444 日分）で放送されたクローズドキャプション用の字幕データを利用した．なお，このデータには，前述の評価データに選んだ番組の字幕も含まれている．

3.3. 語彙の設定

言語モデルの語彙には，NHK 新用字用語辞典や日本語発音アクセント辞典などから選定した基本語彙(101.5K 単語)[4]を使用した．

3.4. サブワード HMM の学習

3.4.1 学習データの収集

サブワード HMM を学習するために，語彙に含まれない単語を学習テキストから抽出し，その読みをカナに変換した．また，学習データを増やすため，学習テキスト中のすべてのカタカナ語を抽出して追加した．この結果得られたカナ表記の単語は，異なり数 7,981，のべ数 133,189 となった．これを音節単位に分割してサブワード HMM の学習に使用した．

3.4.2 HMM の状態数の設定

HMM の状態数を決めるため，最終状態 S=2 の HMM(1 状態出力型)と，S=3(2 状態出力型)の HMM を作成し，それぞれの MDL の値を比較したところ，1 状態出力型による MDL の値の方が小さかった．また，それぞれの HMM から出力される学習データの尤度の和を比べたところ，1 状態出力型による値の方が大きかった．2 状態出力型であれば，単語の始端付近と終端付近に現れやすいサブワードの傾向を各状態で学習できるため，より精度がよいものとなることを期待したが，以上の比較からは，そのような結論が得られなかったことになる．これは，学習データの単語の音節数の平均が 3.2 であったため，サブワード分割に適さず，状態スキップのパスが優先される単語が多かったためと考えられる．さらに，カタカナ語を集めた評価セットを，各 HMM によって最尤サブワード系列に分割する実験を行ったところ，得られた分割パターンはどちらの HMM を用いてもほぼ同じ結果となり，各語の最尤サブワード系列から出力される尤度の和は 1 状態出力型を用いた場合の方が大きかった．以上の結果に加えて，パラメータ数の少なさや，デコーダへの組み込みの容易さも勘案して，今回の実験では 1 状態出力型 HMM(S=2)を採用することにした．

3.4.3 HMM の再推定回数の設定

値が収束するまでパラメータの再推定を繰り返すと，学習データに多いパターンに対して，HMM はより大きな尤度を出力するようになる．一方，学習データに少ないパターンに対しては，きわめて小さな尤度を出力するようになり，再推定の繰り返し回数を増やすほど，この尤度差が増す．あらゆるパターンの未知語を認識できるようにするためには，頑健さを保つ上で，尤度差が大きくなりすぎることは望ましくない．よって今回の実験では，パラメータの再推定回数を 1 回に制限することにした．

3.4.4 サブワード長の上限の設定

サブワード長の上限値 N を大きくするほど，MDL の値は小さくなるが，$N \to \infty$ にすると，得られる HMM は，学習データ自身の単語 1-gram に近づいていく傾向がみられる．この結果，学習データにないパターンに対しては，頑健さがきわめて弱くなってしまう．サブワードはある程度長い方が，音声認識において有利であるが，あらゆる未知のパターンに対する頑健さを保つために，HMM が出力するサブワード長の上限値を最適に定めることが必要となる．そこで，学習データに含まれない別の評価データを用意して，最尤サブワード系列を求めた結果，その系列が出力する尤度の和が $N=5$ のときに極大となることがわかった．この結果から，以降の実験は，$N=5$ として行うことにした．

3.4.5 MDL によるパラメータ数の削減

以上の条件の下で HMM を作成したところ，まず，29,747 個のパターンのサブワードの出力確率が得られた．次に，MDL によってパラメータの削減を行ったところ，サブワード数を 8,209 個に削減できた．

HMM の出力確率のパラメータとして得られたサブワードの音節数に対する頻度の分布を，HMM の学習前(初期モデル)の分布とあわせて，図 2 に示す．パラメータ削減の前後の比較において，1 音節のサブワード数は同一であるが，それ以外については，音節数が多いサブワードの方が，MDL によって選択される割合が大きくなる傾向がみられる．

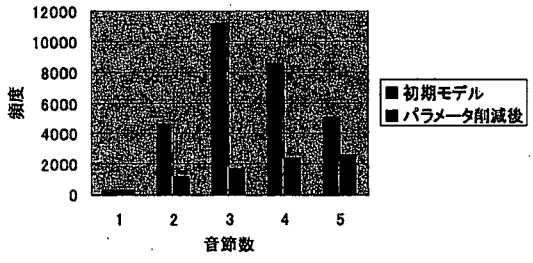

図 2 サブワードの頻度の分布

3.4.5 最尤のサブワード系列を求める予備実験

2.1 節の例に挙げた，「オオアリクイ」「オオスズメバチ」「オオサンショウウオ」「クロサンショウウオ」を，前節で得られた HMM を用いて，最尤サブワード系列に分割する予備実験を行った結果，「オオアリク|イ」「オオ|スズメバチ」「オオサ|ンショウウオ」「クロ|サ|ンショウウオ」と分割された．なお，「クロサンショウウオ」以外は，学習データ中にも存在するものであった．サブワード長(音節数)の上限を $N=5$ としたため，6 音節である「サンショウウオ」「オオアリクイ」などは，必ず分割されることになるが，分割数が少ないサブワード系列の尤度が有利であり，かつ，全般的に 1 音節のサブワードの頻度が大きいこととから，学習データ中に存在する 6 音節の単語は，5 音節と 1 音節に分割されてしまうケースが多く見られた．

3.5. 言語モデル

3.2 節に記した字幕テキストを形態素解析し，表記統一の後処理を行ったもの(1.8M 単語)により，3-gram 言語モデルを作成した．語彙は，3.3 節に記した基本語彙を用いた．評価データに対するパープレキシティーは 28.9，未知語率は 2.0%であった．

3.6. 音響モデル

評価音声には，背景音楽や雑音が多く含まれているため，クリーンな音声で作成した音響モデルでは，良好な認識精度が得られない．そこで，以下の手順によって，評価音声を利用して音響モデルの準教師あり適応[14]を行った．

1. 評価データに対応する字幕テキストだけで言語モデル(3-gram モデル)を作成する．
2. 1 で作成した言語モデルを用いて評価音声の認識を行う．
3. 認識結果を文単位に分割し，文毎に言語モデルの 3-gram ヒット率を求める．
4. ヒット率の高い文を，音声認識が成功した音声区間と判定して，この音声区間を抽出する．
5. 4 で得た音声と対応する字幕テキストの読みを利用して，MLLR と MAP 推定により音響モデルの適応化を行う．
6. 5 で得た音響モデルを用いて，2～5 を繰り返す．

なお，初期の音響モデルとして，大量の NHK ニュースの番組音声で学習した性別非依存状態共有 triphone HMM(16 混合ガウス分布)を利用した．

3.7. デコーダ

音声認識実験に用いたデコーダは，NHK の音声認識装置[15]を改良して，サブワード HMM の認識機能を付加したものである．未知語の認識の際には，式(1)に記したように，未知語クラス(UNK)の N-gram 確率とサブワード HMM の出力確率の積を利用する．

3.8. 認識実験結果

表 1 に，認識実験結果(単語誤り率: WER)を示す．なお，サブワード HMM と比較するための未知語モデルとして，全 1 音節および 5 音節以下の高頻度のサブワードで作成した相対頻度による 1-gram モデルを作成した．なお，このモデルに登録するサブワードのエントリ数を 8,209 として，パラメータ数をサブワード HMM と同一にした．また，本表には，評価音声のうち，未知語を含む文(100 文 1196 単語，未知語率 10.1%)だけでみた場合の WER も記した．

認識結果のうち，複数のサブワードが連続して出力された部分は，1 単語にマージして評価した．また，未知語でない区間にサブワードが出力された場合は，正解単語の読みにマッチしていれば正解とした．例えば，学習テキストにおける標準表記が「おおかみ」であっても，サブワード HMM によって「オオカミ」という表記で出力された場合は正解として扱った．

この結果，未知語率 2.0%の評価音声で，未知語モデルなしの場合に比べて，WER は 15.0%から 14.3%へ 0.7 ポイントの改善がみられた．また，未知語を含む文だけでみると，評価音声の未知語率は 10.1%となるが，WER は 26.7%から 18.4%へ 8.3 ポイントの改善がみられた．これより，それぞれの単語誤り削減率は，5.3%および 31.1%となった．HMM と 1-gram による WER の比較の結果，HMM を用いた場合の方がよい性能が得られることも明らかになった．

表 2 に，未知語区間の単語正解率(Corr.)と，未知語

の前後1単語ずつを加えてみた場合のCorr.を示す．未知語区間における誤りの傾向を調べると，1音節のサブワードの脱落が多くみられた．これは，triphoneを適用できないため，音声認識に不利であるためと考えられる．また，複数のサブワードで認識されるべき単語が不利となり，1つのサブワードで表現される対立候補で誤認識されるケースがみられた(例：N+1音節の「オオアリクイ」がN音節の「コアリクイ」と誤認識された)．さらに，未知語区間を正しく検出できても，サブワード間の制約がないため，正しいサブワード系列が出力されないケースもみられた．今回の評価音声は，音響的な環境が劣悪であるため，これについては，音響モデルの性能向上や，単語間triphoneの導入などによって解決できる可能性がある．

表1 評価音声の認識実験結果 (WER %)

未知語モデル	評価文全体	未知語を含む文
なし	15.0	26.7
1-gram	14.5	20.4
HMM	14.3	18.4

表2 未知語区間の単語正解率 (Corr. %)

未知語モデル	未知語	未知語+前後1単語
なし	0.0	53.2
1-gram	41.5	72.3
HMM	48.3	76.0

4. まとめ

あらゆるパターンの未知語をカナ文字列によって受理可能にするため，カナ表記の単語に含まれる1音節を最小単位とした任意のサブワードを未知語の認識単位に定め，このサブワードからなる可変長のシンボル系列を出力するHMMを構築した．つづいて，MDL規準によるモデルのパラメータの削減を行った．これによって得られたHMMを用いて音声認識実験を行ったところ，単語誤り削減率が5.3%であった．これを，未知語を含む文だけで観察した場合の誤り削減率は31.1%であった．また，今回作成したHMMと同一のパラメータ数をもつサブワード1-gramモデルの比較では，HMMを用いる方が良好な認識精度が得られることを確認した．

問題として，未知語の表現のために多くのサブワードを必要とする場合と，サブワードの音節数が少ない場合(特に1音節)に，音声認識において不利であることが明らかになった．よって今後は，1つの未知語を表現するのに要するサブワード数と各サブワードの音節数に応じて確率値を制御する手法や，単語間triphoneの導入などについて検討したい．また，本手法では学習データと評価データの類似性で性能が左右されると考えられるため，別タスクにおける本手法の検証や，サブワードHMMの学習データをさらに増やした場合の検証も行いたい．学習データを増やすと，パラメータとなるサブワード数の爆発が予想されるため，より効果的なサブワードの選択手法の検討も必要であると考えられる．

文　献

[1] 安藤彰男，今井亨，小林彰夫，本間真一，後藤淳，清山信正，三島剛，小早川健，佐藤庄衛，尾上和穂，世木寛之，今井篤，松井淳，中村章，田中英輝，都木徹，宮坂栄一，磯野春雄，“音声認識を利用した放送用ニュース字幕制作システム，”信学論(D)，vol.J83-D2，no.6，pp.877-887，Jun.2001.

[2] 松井淳，本間真一，小早川健，尾上和穂，佐藤庄衛，今井亨，“言い換えを利用したリスピーク方式によるスポーツ中継のリアルタイム字幕制作，”信学論(D)，vol.J87-D2，no.2，pp.427-435，Feb.2004.

[3] 佐藤庄衛，小林彰夫，尾上和穂，山田一郎，佐野雅規，今井亨，“メタデータ生成ための音声認識の改善，”映メ年次大会，9-1，Aug.2005.

[4] 本間真一，小林彰夫，尾上和穂，佐藤庄衛，今井亨，“情報番組のための基本語彙と想定重要語彙を利用した音声認識，”音講論，2-1-5，pp.79-80，Mar.2005.

[5] 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，SP91-06，pp.41-47，Dec.1991.

[6] 甲斐充彦，廣瀬良文，中川聖一，“単語N-gram言語モデルを用いた音声認識システムにおける未知語・冗長語の処理，”情処論，vol.40，no.4，pp.1383-1394，Apr.1999.

[7] 内山将夫，松本宏，“仮名文字と連語登録を併用した統計的言語モデル，”信学技報，SP99-38，pp.87-94，Jun.1999.

[8] 廣瀬良文，伊藤克亘，鹿野清宏，中村哲，“日本語ディクテーションシステムにおける被覆率の高い言語モデル，”信学論(D)，vol.J83-D2，no.11，pp.2300-2308，Nov.2000.

[9] M. Bisani and H. Ney, “Open vocabulary speech recognition with flat hybrid models,” Proc. INTERSPEECH 2005, pp.725-728, Sep.2005.

[10] 山本博史，小窪浩明，菊井玄一郎，小川良彦，匂坂芳典，“複数のマルコフモデルを用いた階層化言語モデルによる未登録語認識，”信学論(D)，vol.J87-D2，no.12，pp.2014-2111，Dec.2004.

[11] 服部多栄子，椎名努，堂免大規，“生字幕放送サービスーシステムとサービスの概要ー，”映メ年次技報，Vol.28，no.5，pp.17-20，Jan.2004.

[12] J. Rissanen, “Universal coding, information, prediction, and estimation,” IEEE Trans. IT, vol.30, no.4, pp.629-636, Jul.1984.

[13] S. Deligne and F. Bimbot, “Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams,” Proc. IEEE International Conf. on Acoustics, Speech, and Signal Processing, pp.169-172, May 1995.

[14] L. Lamel, J.-L. Gauvain, and G. Adda, “Lightly supervised and unsupervised acoustic model training,” Computer Speech and Language, vol.16, pp.115-129, Jan.2002.

[15] T. Imai, A. Kobayashi, S. Sato, S. Homma, K. Onoe, and T. S. Kobayakawa, “Speech recognition for subtitling live broadcast,” Proc. ICA2004, vol.1 pp.165-168, Apr.2004.