Vol.2010-SLP-82 No.15
2010/7/24

未知語認識のための仮名・漢字単位の構築手法と性能評価

久保慶伍[†1]　川波弘道[†1]
猿渡洋[†1]　鹿野清宏[†1]

音声認識の実用化が進み，Voice Search や音声ドキュメント検索のような幅広いドメインを扱うタスクが増えてきた．幅広いドメインを扱うタスクでは，ドメインを限定したタスクと比べて，辞書に含まれない単語「未知語」が高い頻度で出現する．特に，日本語や中国語のような言語は，漢字により複合語や造語，省略語などの未知語が生まれやすいため未知語の対処が重要な課題となる．未知語の対処法として，サブワードと呼ばれる形態素よりも短い単位を形態素単位の認識辞書と言語モデルに組み込むことにより，未知語を表現する手法がある．その他，Google の音声検索に見られるようにコーパスを増やして，ヒットレートを上げることも未知語の対策には有望なアプローチである．本報告では漢字で表現される未知語に対して頑健な音声認識を実現するために，仮名・漢字単位を構築し，構築した仮名・漢字単位をサブワードとして採用する．評価実験では（1）仮名・漢字単位と形態素単位の混合 N-gram と，従来手法として (2) 形態素単位 N-gram と（3）音節列の統計的特徴を用いて決定した音節・音節列単位と形態素単位の混合 N-gram の 3 つの言語モデルを構築し，日本の地名に関する検索発話のテストセットで評価を行った．評価実験の結果から，仮名・漢字単位と形態素単位の混合 N-gram は 1-best の文正解率において，形態素単位 N-gram よりも 6.61%高く，統計的に音節列を 200 種類採用した音節・音節列単位と形態素単位の混合 N-gram よりも 0.83%高い値を示した．

Building Method and Evaluation of Kana and Kanji Unit for Recognition of OOV

Keigo KUBO,[†1] Hiromichi KAWANAMI,[†1]
Hiroshi SARUWATARI[†1] and Kiyohiro SHIKANO[†1]

Following the success of real-time ASR, voice search and spoken document retrieval services have been put into practice. These services are expected to deal with wide range vocabulary including OOV (Out-of-vocabulary) words to the system such as new words or persons' names. To treat with OOV problem, a unit called sub-word, which is shorter than a morpheme are introduced to speech recognition. In this paper, the authors construct a set of sub-words consist of Kana and Kanji units. The set is introduced to construct a N-gram language model with conventional morpheme units. For comparison, a morpheme N-gram model and a N-gram consists of morphemes with selected phoneme sequence are also constructed. The experiments are conducted using Japanese place name utterances as a test set. As the result, the sentence accuracy of 1-Best result of the proposed method is 6.61% and 0.83% higher than the conventional morpheme N-gram model and the morpheme with selected phoneme sequence N-gram model, respectively.

1. はじめに

音声認識の実用化が進み，Voice Search や音声ドキュメント検索のような幅広いドメインを扱うタスクが増えてきた．Voice Search の展望としては大量のデータによるヒットレートの向上が重要となると考えられるが，幅広いドメインを扱うタスクでは日に日に多種多様な新しい語彙が出現するため，それらを全て辞書に登録することは事実上不可能である．このような辞書に登録されていない単語を「未知語」といい，幅広いドメインを扱うタスクにおいては狭いドメインよりも未知語が出現しやすいため，未知語の対処法が重要となる．

未知語の対処法として，サブワードと呼ばれる形態素よりも短い単位を形態素単位の認識辞書と言語モデルに組み込むことにより，未知語を表現する手法がある[1)2)]．この手法において，音素といった短いサブワードを用いると言語的制約が弱くなり、認識性能が低下する．その反対に，長いサブワードを用いると多種多様な未知語を表現することができなくなる．そのため，多種多様な未知語を表現する表現力を維持しながら言語的制約を強めるために，音節列の統計的特徴を用いて音節列を決定し，音節とその音節列（以後，統計的音節・音節列単位）をサブワードとして採用する．しかし，統計的音節・音節列単位はヒューリスティックに採用する音節列の数を決める必要がある．その他，Google の音声検索に見られるようにコーパスを増やして，ヒットレートを上げることも未知語の対策には有望なアプローチである．

本報告では漢字で表現される未知語に対して頑健な音声認識を実現するために，仮名・漢字単位を構築し，構築した仮名・漢字単位をサブワードとして採用する．日本語や中国語のような言語は，漢字により複合語や造語，省略語などの未知語が生まれやすい．仮名・漢

†1 奈良先端科学技術大学院大学 情報科学研究科
Graduate School of Information Science, Nara Institute of science and Technology

Vol.2010-SLP-82 No.15
2010/7/24

字単位はそのような未知語を頑健に認識できると考えられる．また，音節列をヒューリスティックに決める必要がないという利点がある．

この仮名・漢字単位をサブワードとして採用した手法を評価するため，評価実験では（1）仮名・漢字単位と形態素単位の混合 N-gram と，従来手法として (2) 形態素単位 N-gram と（3）音節列の統計的特徴を用いて決定した音節・音節列単位と形態素単位の混合 N-gram の 3 つの言語モデルを構築した．そして，日本の地名に関する検索発話のテストセットで評価を行った．

2. 統計的音節・音節列単位

統計的音節・音節列単位をサブワードとして用いる場合，音節列の採用にエントロピーを用いた手法[1] や，頻度を用いた手法[2] がある．しかし，最適な音節列の採用数は扱うドメインの大きさや種類により異なると考えられるため，これらの手法は音節列の採用数や音節列を採用する頻度をヒューリスティックに決めなければならない．本報告では仮名・漢字単位の性能評価のために統計的音節・音節列単位 N-gram を構築し，その性能を評価した．

2.1 統計的音節・音節列単位 N-gram の構築手順

統計的音節・音節列単位 N-gram の構築手順を図 6 に示す．まず，語彙の発音を特定し，その発音を文字に分割する．次に採用する音節列を決定・連結し，それにより得られるコーパスから統計的音節・音節列単位 N-gram を構築する．採用する音節列は条件付きエントロピーにより決定した．

2.2 採用する音節列の決定・連結

条件付きエントロピーを用いて音節列を決定・連結し，統計的音節・音節列用の学習コーパスを構築する手順を以下に示す．

(1) 図 1 の手順 2 より得られる音節単位に区切られたコーパスをコーパス C とする．また，説明の簡便さのために音節または音節列をユニットと定義する．

(2) コーパス C に出現する連続する 2 個のユニット全てに対し，(3) と (4) を行う．

(3) コーパス C に出現する連続する 2 個のユニットを連結して一つのユニットとして扱い，ユニットを連結したコーパスをコーパス $\hat{C}$ とする．

(4) コーパス $\hat{C}$ の条件付きエントロピーを以下の式により求める．

$$H(X|Y) = -\sum_{y} Pr(Y=y) \sum_{x} Pr(X=x|Y=y) log Pr(X=x|Y=y) \quad (1)$$

X は学習コーパスに出現する任意のユニットの後ろに出現するユニットを表す確率変数である．Y は学習コーパスに出現する任意のユニットの前に出現するユニットを表す確率変数である．$Pr(Y=y)$ はユニット y の出現確率である．$Pr(X=x|Y=y)$ はユニット y の後にユニット x が出現する確率である．$Pr(Y=y)$，$Pr(X=x|Y=y)$ の値は共にコーパス $\hat{C}$ から最尤推定で求める．

(5) (2)〜(4) で求めた条件付きエントロピーの内，最も条件付きエントロピーが低かったコーパス $\hat{C}$ を次のコーパス C とする．

(6) 音節列の採用数だけ (2)〜(5) を繰り返す．

この手順により連接関係の強い音節列が統計的に決定・連結され，統計的音節・音節列用の学習コーパスが得られる．

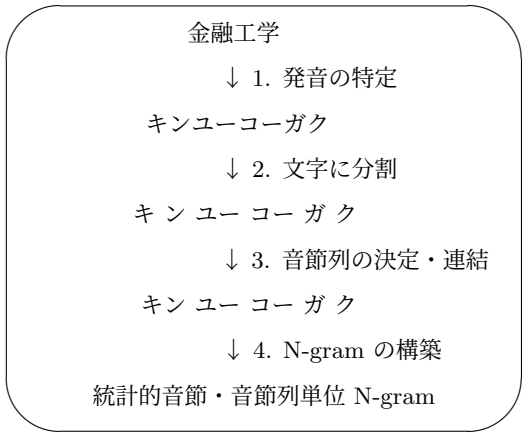

図 1 統計的音節・音節列単位 N-gram - 構築手順の例
Fig. 1 statistical syllable and syllable row unit N-gram - example of build process

3. 仮名・漢字単位

仮名・漢字単位 N-gram を用いた先行研究として，山田らが仮名・漢字単位で N-gram を構築する手法を提案している[3]．これは，文を構成する各文字に対して発音の情報が付与されている仮名・漢字用のコーパスから N-gram モデルを構築する手法である．この手法は漢字 1 文字の発音を音節列として採用しているため，統計的音節・音節列単位のようにヒューリスティックに音節列の採用数を決める必要がない．

しかし，仮名・漢字用のコーパスを用意するために，文字と発音の対応をとる必要があ

Vol.2010-SLP-82 No.15
2010/7/24

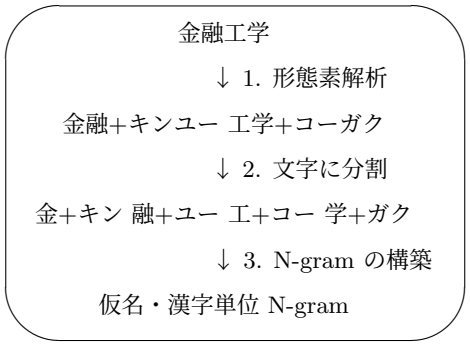

図 2 仮名・漢字単位 N-gram - 構築手順の例
Fig. 2 Kana and Kanji unit N-gram - example of build process

る．筆者らはこの問題に対して発音割り当て手法を提案している[4]．本報告でもその手法を用いて仮名・漢字単位 N-gram を構築する．

3.1 仮名・漢字単位 N-gram の構築手順

仮名・漢字単位 N-gram の構築手順を具体例とともに図 2 に示す．まずテキストコーパスの形態素解析を行う．次に，各形態素を文字に分割し，各文字に対応する発音を付与する．そして，各文字に発音を付与したコーパスから仮名・漢字単位 N-gram を構築する．構築の際に「香具師 (ヤシ)」などの表記の文字数に対して読みの文字数が少ないものは分割せずにその形態素を 1 文字として扱かった．

図 2 の手順 2 では各文字に対してどの発音が割り当たるのか求める必要がある．しかし「工学+コーガク」から「工+コー 学+ガク」を求める際,「工」は「コー」と発音するのか「コーガ」と発音するのか，この情報だけでは判断できない．そこで，漢字 1 文字の発音が集録された辞書が必要となるが、一般の辞書には促音化や濁音化など，漢字の組み合わせによって特殊な発音になる場合の漢字 1 文字の発音は集録されていない．このような特殊な読みは地名や人名の読みによく見られる。そこで，漢字を含んだ文字列と発音の対データを大量に用意し，そのデータを用いて，反復法である発音割り当て手法により，ある文字にどの発音が割り当たるかを求め、仮名・漢字用のコーパスを自動で構築する．そして，そのコーパスから仮名・漢字単位 N-gram を構築する．

3.2 発音割り当て手法

発音割り当て手法を以下に示す．

(1) 任意の辞書から表 1 のような発音と漢字の対データを作成する．

表 1 発音と漢字の対データ
Table 1 Pair data of pronunciation and kanji

コーガク 工学
コーギョー 工業
ミニクイ 見にくい
アジワイブカイ 味わい深い

表 2 各文字に対する発音の割り当ての組み合わせ
Table 2 Combination of allocation of pronunciation to each character

工+コー 学+ガク
工+コーガ 学+ク

(2) (3) と (4) を全ての対データに対して行う．

(3) 対データから表 2 のように各文字に対する発音の割り当ての組み合わせを全て展開する．展開した組み合わせを発音仮説と定義する．

(4) 展開した組み合わせから，その組み合わせに出現する全ての文字とそれに割当たる読みに対して，以下の式を求める．

$$P(S|K,l_K) = \frac{\max\limits_{\{i|K\Rightarrow S\}} \prod\limits_{j=1}^{N_{l_K}} P(s_{ij}|k_j)}{\sum\limits_{S_m\in\{S_1,\ldots,S_M\}_{l_K}} \max\limits_{\{i|K\Rightarrow S_m\}} \prod\limits_{j=1}^{N_{l_K}} P(s_{ij}|k_j)} \tag{2}$$

$P(S|K,l_K)$ は文字 K が l 番目に出現した対データにおいて，文字 K に発音 S が割り当てられる確率である．$\{i|K\Rightarrow S\}$ は文字 K に対して発音 S が割り当てられている発音仮説の集合である．N_{l_K} はこの対データの表記の文字数である．$P(s_{ij}|k_j)$ は反復法により更新するパラメータであり，対データの表記の j 番目の文字 k_j に対して，発音 s_{ij} が割り当てられる確率を表している．ただし，初期値は $P(s_{ij}|k_j)=1$ とする．$\{S_1,\ldots,S_M\}_{l_K}$ はこの対データの文字 K に割り当てられている発音の種類である．

(5) 文字 K に対して発音 S が割り当てられる確率 $P(S|K)$ を以下の式により更新する．

$$P(S|K) = \frac{\sum\limits_{l_K=1}^{L_K} P(S|K,l_K)}{L_K} \tag{3}$$

L_K は文字 K が出現した対データの総数である．

(6) (2)〜(5) を任意回数繰り返す．

(7) 発音が付与された形態素から，発音の割り当ての組み合わせを全て展開し，発音仮説

を求める．そして，更新した確率 $P(S|K)$ を用いて，以下の式を最大にする発音仮説 $\hat{i}$ を求め，発音仮説 $\hat{i}$ の発音割り当てを採用する．

$$\hat{i} = \arg\max_{i} \sum_{j=1}^{N} \log P(s_{ij}|k_j) \quad (4)$$

これにより，各文字ごとに発音が付与された仮名・漢字用のコーパスを得ることができる

4. 評価実験

この仮名・漢字単位をサブワードとして採用した手法を評価するため，評価実験では（1）仮名・漢字単位と形態素単位の混合 N-gram と，従来手法として (2) 形態素単位 N-gram と（3）音節列の統計的特徴を用いて決定した音節・音節列単位と形態素単位の混合 N-gram の 3 つの言語モデルを構築した．そして，日本の地名に関する検索発話のテストセットで評価を行った．

4.1 実験条件

実験条件を表 3 に示す.

4.1.1 学習コーパス

形態素用の学習コーパスには全国住所リストと全国駅名リストを使用した．全国住所リストは郵便番号データから住所部分を抽出し，表 4 のように都道府県，市，区，町，丁目で検索発話として考えられる組み合わせを展開したコーパスである．また，全国駅名リストは全国駅名一覧から表 5 のように「駅」を付与する文と付与しない文の 2 つに展開したコーパスである．

また，統計的音節・音節列用の学習コーパスと仮名・漢字用の学習コーパスの一部を表 6, 表 7 に示す．これらの学習コーパスは郵便番号データと全国駅名一覧から抽出した語彙である．<s> と </s> はそれぞれ開始記号と終了記号である．

表 6 の統計的音節・音節列用の学習コーパスは節 2.2 の手順により構築したコーパスである．音節列の採用数は 200 である．

表 7 の仮名・漢字用の学習コーパスは節 3.2 の発音割り当て手法により構築したコーパスである．発音割り当て手法の反復回数は 4 回である．表 7 から，節 3.2 の発音割り当て手法により各文字に対して正しい読みが付与されていることがわかる．ただし，仮名・漢字単位 N-gram を構築する際には仮名・漢字用の学習コーパスの表記部分を削除する処理を行った．つまり,「<s> 下+シ モ+モ 川+カワ 原+ラ </s>」を「<s> シ モ カワ ラ </s>」

表 3 実験条件
Table 3 Experiment condition

学習コーパス	形態素用	全国住所リスト（郵便番号データ*1から住所部分を抽出し，都道府県・市・区・町・丁目で検索発話として考えられる組み合わせを展開，約 47 万文) + 全国駅名リスト (全国駅一覧*2から「駅」ありの文となしの文の 2 つに展開，約 1 万 8 千文)
	統計的音節・音節列用 仮名・漢字用	郵便番号データ，全国駅名一覧から抽出した語彙
テストセット	地名に関する検索発話（男性 36 人，女性 2 人が各々考えた地域に関する検索発話，484 発話）	
	学習コーパスに含まれている地名に関する検索発話（417 発話）	
形態素解析器	chasen 2.4.4	
形態素辞書	ipadic-2.7.0 郵便番号データ，全国駅名一覧から抽出した語彙の辞書	
音響モデル	JNAS PTM モデル	
認識エンジン	Julius 4.1.4	
認識方法	2pass 認識 （前向き 2-gram, 後ろ向き 3-gram）	形態素単位 N-gram（形態素） 形態素単位 + 統計的音節・音節列単位 50 N-gram（音節列 50） 形態素単位 + 統計的音節・音節列単位 100 N-gram（音節列 100） 形態素単位 + 統計的音節・音節列単位 150 N-gram（音節列 150） 形態素単位 + 統計的音節・音節列単位 200 N-gram（音節列 200） 形態素単位 + 仮名・漢字単位 N-gram（仮名・漢字）
	孤立単語認識	文に読みを付与したものを単語として認識
融合重み	形態素単位が 0.9341，統計的音節・音節列単位または仮名・漢字単位が 0.0659	
評価方法	1-best, 5-best, 10-best の文正解率（仮名で評価）	

になるよう処理を行った．これは，表記を与えることにより未知語を認識できないほど強力な言語的制約が掛かることを防ぐためである．

4.1.2 テストセット

テストセットは男性 36 人、女性 2 人が各々考えた日本の地域に関する検索発話である．このテストセットは合計で 484 発話ある．表 8 にテストセットの一部を示す.

表 9 において今回のテストセットで未知語となる語彙の一部を示す．未知語の種類として「市」などを省略した省略語と住所や駅名以外の名詞がある．表 10 にヒットレートを示す. 1-gram ヒットレートは 93.41%，2-gram ヒットレートは 86.02%, 3-gram ヒットレートは 84.66%, sentence ヒットレートは 79.75%である．

また，未知語が存在しない発話において仮名・漢字単位をサブワードとして採用する際の認識性能について調べるために，地名に関する検索発話から学習コーパスに含まれている発話を抽出したテストセットを用意した．

Vol.2010-SLP-82 No.15
2010/7/24

表 4 全国住所リスト
Table 4 Across-the-country address list

札幌市
中央区
北海道札幌市
札幌市大通西
北海道札幌市大通西
北海道
札幌市中央区
北海道札幌市中央区
札幌市中央区大通西
北海道札幌市中央区大通西
大通西
北海道大通西
大通西 1 丁目
札幌市大通西 1 丁目
札幌市中央区大通西 1 丁目
北海道大通西 1 丁目
北海道札幌市大通西 1 丁目

表 5 全国駅名リスト
Table 5 Across-the-country station list

相生
相生駅
相賀
相賀駅
秋鹿町
秋鹿町駅
合川
合川駅
相川
相川駅
愛環梅坪
愛環梅坪駅
愛甲石田
愛甲石田駅
愛山
愛山駅

表 6 統計的音節・音節列用の学習コーパス
Table 6 Training corpus for statistical syllable and syllable row unit

<s> シモ カワ ラ </s>
<s> ナ ミ オカ キ チ ナイ </s>
<s> ナカ ツ クマ </s>
<s> ジョウ ハナ </s>
<s> ソ ト オカ </s>
<s> カ ミナ ガ イ </s>

表 7 仮名・漢字用の学習コーパス
Table 7 Training corpus for Kana and Kanji unit

<s> 下+シ モ+モ 川+カワ 原+ラ </s>
<s> 浪+ナミ 岡+オカ 吉+キチ 内+ナイ </s>
<s> 中+ナカ 津+ツ 隈+クマ </s>
<s> 城+ジョウ 端+ハナ </s>
<s> 外+ソト 岡+オカ </s>
<s> 上+カミ 永+ナガ 井+イ </s>

表 8 日本の地域に関する検索発話
Table 8 Retrieval utterance concerning region in Japan

島根県
大阪府泉大津市
鹿児島郡三島村
鶴橋
屋久島
九十九里浜
デンデンタウン

表 9 テストセットに出現する未知語
Table 9 OOV that appears in test set

稚内
デンデンタウン
新世界
屋久島
明日香村
九州
奥飛騨

表 10 テストセットのヒットレート
Table 10 Hit-rate of test set

1-gram hit-rate	93.41%
2-gram hit-rate	86.02%
3-gram hit-rate	84.66%
sentence hit-rate	79.75%

表 11 実験結果 - 地名に関する検索発話（仮名による文正解率，手法の表記は表 3 に従う.）
Table 11 Experiment result - retrieval utterance concerning region in Japan (sentence correct with Kana, Notations of methods follow Table 3.)

		地名に関する検索発話（484 発話）		
認識手法		1-best	5-best	10-best
音節列	50	73.55%	83.47%	85.74%
	100	74.17%	83.68%	85.95%
	150	75.21%	83.68%	85.95%
	200	75.41%	83.88%	85.95%
仮名・漢字		76.24%	84.30%	85.95%
形態素		69.63%	75.41%	76.65%
孤立単語認識		72.11%	76.03%	76.86%

4.1.3 認識方法

評価に用いた認識方法は前向き 2-gram と後ろ向き 3-gram の 2pass による音声認識（以後，2pass 認識）と孤立単語認識である．2pass 認識では提案手法である（1）仮名・漢字単位 N-gram と形態素単位 N-gram の混合 N-gram（実験結果では仮名・漢字と表記）と従来手法である（2）形態素単位 N-gram（実験結果では形態素と表記）と（3）音節列を節 2.2 の手法により 50,100,150 または 200 種類採用した統計的音節・音節列単位 N-gram と形態素単位 N-gram の混合 N-gram（実験結果では音節列 50,100,150,200 と表記）の 3 つを評価した．今回の評価実験ではテストセットにおける未知語率を既知とし，混合重みは表 10 の 1-gram ヒットレートから，形態素単位 N-gram を 0.9341，統計的音節・音節列単位 N-gram または仮名・漢字単位 N-gram を 0.0659 にした．また，孤立単語認識は学習コーパス内の文に読みを付与したものを単語として認識を行った．

4.1.4 評価方法

1-best, 5-best, 10-best の仮名における文正解率で評価した．

4.2 実験結果と考察

実験結果を表 11，表 12 に示す．表 11 から，地名に関する検索発話における文正解率は

表 12 実験結果 - 学習コーパスに含まれている地名に関する検索発話（仮名による文正解率，手法の表記は表 3 に従う.)
Table 12 Experiment result - retrieval utterance concerning region in Japan included in training corpus (sentence correct with Kana, Notations of methods follow Table 3.)

		学習コーパスに含まれている地名に関する検索発話（417 発話）		
認識手法		1-best	5-best	10-best
音節列	50	80.10%	87.77%	89.69%
	100	80.82%	88.01%	89.69%
	150	81.77%	87.77%	89.69%
	200	81.77%	87.77%	89.45%
仮名・漢字		82.49%	88.25%	88.97%
形態素		82.97%	88.73%	89.93%
孤立単語認識		81.53%	86.09%	87.05%

形態素よりも 1-best, 5-best, 10-best 共に音節列 50,100,150,200 と仮名・漢字の方が高い．これは「屋久島」や「稚内」などの未知語を統計的音節・音節列単位や仮名・漢字単位を形態素単位に組み入れることで認識可能になったからである．また，仮名・漢字と音節列 200 を比較すると 1-best では 0.83%，5-best では 0.52%ほど仮名・漢字の文正解率が高かった．これは，仮名・漢字単位の方が採用する音節列が多く，言語的制約が強いからだと考えられる．統計的音節・音節列単位も音節列の採用数を上げれば言語的制約が強くなり文正解率が高くなることが表 11 から推測できる．しかし，採用する数が増加するほど，長い音節列が採用され，未知語を表現する能力が低下すると考えられる．そのため，統計的音節・音節列単位では最適な音節列の採用数が，扱うドメインの大きさや種類により異なると考えられるため，音節列の採用数をヒューリスティックに決めなければならない．一方で仮名・漢字単位は漢字一文字の発音を音節列に採用するため，ヒューリスティックに音節列の採用数を決める必要がない．しかも，漢字単位で音節列を採用しているため，漢字の組み合わせにより構成される未知語に対して，その未知語を表現する能力が失われないという利点がある．

表 12 から，学習コーパスに含まれている地名に関する検索発話の形態素と仮名・漢字の文正解率を比較すると，1-best，5-best では 0.5%ほど形態素が高かった．また，仮名・漢字と音節列 200 の文正解率を比較すると，1-best では 0.72%，5-best では 0.48%ほど仮名・漢字が高かった．この実験結果から，仮名・漢字単位をサブワードとして形態素単位に組み込むと形態素単位よりも認識候補が増えるため，文正解率に若干の悪影響を与えていることがわかる．しかし，統計的音節・音節列単位と比べると 5-best までならその影響は少ないことがわかる．これも，統計的音節・音節列単位より仮名・漢字単位の方が言語的制約が強いからだと考えられる．

また，表 11 と表 12 の 10-best から，10-best まで考慮すると仮名・漢字単位と統計的音節・音節列単位の差はほとんど無いと考えられる．

5. おわりに

本報告では漢字で表現される未知語に対して頑健な音声認識を実現するために，仮名・漢字単位を構築し，構築した仮名・漢字単位をサブワードとして形態素単位の認識辞書と言語モデルに組み込み，未知語の音声認識を試みた．

実験結果から，未知語を含むテストセットにおける文正解率は形態素単位よりも統計的音節・音節列単位や仮名・漢字単位の方が高かった．また，10-best まで考慮すると統計的音節・音節列単位と仮名・漢字単位の差はほとんど存在しないものの，5-best までは仮名・漢字単位の有効性がみられた．この結果から，ヒューリスティックに音節列の採用数を決める必要がない仮名・漢字単位は漢字の組み合わせにより構成される未知語に対して有効であると考えられる．

今回の評価実験ではテストセットとして日本の地域名に関する検索発話を用いたが，日本の地域名に関しては WEB に点在する地域名に関するデータベースにより，ほぼ網羅できると考えられ，未知語対策を行うメリットは少ない．そのため，今後は網羅することが困難な人名に関してテストセットを構築し，評価したいと考えている．

6. 謝　辞

本研究の一部は，戦略的創造研究推進事業「共生社会に向けた人間調和型情報技術の構築」(JST/CREST) の援助を受けて行われた．

参考文献

1) K. Tanigaki et al, “A hierarchical language model incorporating class-dependent word models for OOV words recognition,” In ICSLP-2000, vol. 3, pp. 123-126.
2) 山本 他, “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,” 電子情報通信学会論文誌, Vol. J87-D-II, No. 12, pp. 2104-2111, 2004.
3) 山田 他, “音声認識における仮名・漢字文字連鎖確率に基づく統計的言語モデルの利用,” 電子情報通信学会論文誌, vol. J77-A , No. 2, pp. 198-205, 1994.
4) 久保 他, “Voice Search のための文字単位 Ngram の検討,” 日本音響学会講演論文集, 3-6-9, pp. 109-112, March 2010.