A new language model is proposed to cope with the demands for recognizing out-of-vocabulary (OOV) words not registered in the lexicon. This language model is a class N-gram incorporating a set of word models that reflect the statistical characteristics of the phonotactics, which depend on the lexical classes. Utilization of class-dependency enhances recognition accuracy and enables identification of the class of OOV words. OOV words can be recognized as transcribed portions having class labels, which provide semantic attributes of OOV words to subsequent language processing. Experimental application of the model to Japanese personal and family names showed that it performs nearly as well as the upper bound of the in-vocabulary recognition.

1. INTRODUCTION

Progress in speech recognition has led to the development of large-vocabulary continuous speech recognition (LVCSR) [1]. However, the LVCSR paradigm does not completely solve the problems presented by out-of-vocabulary (OOV) words. Particularly in languages such as Japanese and Chinese, OOV compound/coined/abbreviated words pose serious problems due to the high word-building capacity of Chinese characters.

One way to cope with these problems is to use a confidence measure based on the likelihood ratio between the in-vocabulary recognition score and a purely acoustic score outside the lexical constraints [3][4]. This measure is used to verify whether an utterance was correctly decoded to in-vocabulary words or it included OOV words. However, this approach has drawbacks in that it requires another procedure to calculate the comparative score, as well as requires other criteria to determine the threshold.

Another approach uses a set of subwords called “filler fragments” [2]. These subwords are imported as pseudo-words to both the lexicon and the N-gram language model and are used to fill (transcribe) the OOV regions among in-vocabulary words. This approach can protect the segmentations on in-vocabulary words damaged by neighboring OOV words, but the model does not sufficiently represent the linguistic characteristics of/around OOV words.

This paper describes a hierarchical language model to cope with the OOV problems. The model exploits the statistical characteristics of the concatenations of phonetic units, which are usually peculiar to lexical classes (e.g., person-name and company-name). Utilization of class-dependency enhances the accuracy of the model and also enables the identification of the class of OOV word. OOV words are recognized as portions of transcriptions with the class labels, which provide semantic attributes of OOV words, such as proper nouns, to subsequent language processing.

2. PHONOTACTIC FEATURES OF LEXICAL CLASSES

Our study on OOV words started with the two classes: Japanese family names and personal names. The objective is to correctly identify the readings and the class of the OOV words in continuous speech recognition. To identify the OOV words,
it is empirically prospected being effective to employ the statistics of the following phonotactics.

First, there are frequently adopted subword units (short concatenations of phonemes) in the readings of words, resulting from the construction of a Japanese word as a sequence of Chinese characters. For example, the Chinese characters of “tani” (meaning “valley”), “yama” (“mountain”), and “saka” (“hill”) in the authors’ names are frequently occurring units in Japanese family names.

Second, word length consisting of three or four morae per name occurs the most, and again, all three of the authors’ have four morae in both personal and family names. Figure 1 shows the word length (measured by the number of morae) against the number of words. We analyzed a list of 300,000 Japanese names [1]. Note that, in this paper, we assume the same probability for the occurrence of individual names; each name on the list is considered to occur at a frequency proportional to the number of people with that name. The statistics were compared against one million words in a travel arrangement corpus [6], excluding 6,000 Japanese names. The sizes of the corpora are shown in table 1. In figure 1, we can see that Japanese personal and family names take sharp distributions in their length. The names with over three or four morae covered approximately 90% of the names on the list.

3. THE LANGUAGE MODEL

The proposed language model is a class N-gram having a hierarchical structure. Figure 2 depicts an example of the language model used to produce the observation “... name is Koichi Tanigaki”, where the words Koichi and Tanigaki are OOV words of a Japanese Personal Name (JPN) and a Japanese Family Name (JFN). These words are recognized as transcriptions with the class labels, which provides semantic attributes of the OOV words to subsequent language processing.

The upper layer of the model is an inter-class N-gram, which accounts for the rather grammatical aspect of the occurrence of an OOV according to its context. The lower layer consists of a set of intra-class models either of two types: (A) word 1-grams of vocabulary (as in conventional class N-grams, omitted from figure 2) or (B) stochastic word models of the classes where OOV words are predicted to occur (i.e., Japanese family and personal names). The word model has a structure to produce the subword sequence of an OOV word at the appropriate probability according to the phonotactic constraints of the class. The details and formalization of the language model are described in the following subsections.

3.1. Modeling of Words Including OOVs

Let $W$ be word sequence $w_1,w_2,…,w_i$ and $c_i$ be the class of word $w_i$, a class N-gram evaluates the linguistic likelihood of the word sequence $W$ as

$$p(W) = \prod_i p(w_i | w_{i-1},…,w_1)$$

$$= \prod_i p(w_i | c_i)p(c_i | c_{i+1},…,c_1)$$

(1)

Note that our $w_i$ is sometimes an OOV word. Suppose that the class of an OOV can always be found in the existing classes and that the class is known (even after decoding). The probability $p(c_i | c_{i+1},…,c_1)$ then does not change even for OOV $w_i$, and we get equation (1) by estimating the probability of $p(w_i | c_i)$. If we treat OOV $w_i$ as unknown phone sequence $R_i$ (readings of $w_i$ not registered in the lexicon), we can
replace probability $p(w_i | c_i)$ as follows:

$$p(w_i | c_i) = p(R_i | c_i)$$

We call $p(R | c)$ a word model: the phonotactic constraints dependent on a class $c$. The next subsection describes the modeling of this probability.

### 3.2. Phonotactic Modeling of OOV

To get a good estimation of $p(R | c)$—the probability in which readings $R$ of a certain OOV word occurs in class $c$—it is most effective to put appropriate structure of the class. As we described in section 2, the target classes in this paper are Japanese personal and family names, and their readings have statistically salient features, particularly their length and subword units. Thus, we decompose $p(R | c)$ as in equation (5), where $lenR$ denotes the length of phone sequence $R$.

$$p(R | c) = p(lenR | c)p(R | c, lenR)$$

In equation (5), $p(lenR | c)$ denotes the duration constraints on word length. This probability is modeled by a class-dependent Gamma distribution on the number of morae per word. The $p(R | c, lenR)$ denotes the probability at which a phone sequence of length $lenR$ actually becomes $R$ according to class $c$. This probability is modeled by a subword N-gram. Note that the N-gram includes no transition to the word-ending symbol because we have the model of $p(lenR | c)$. The subword units of the N-gram are composed of single and concatenated morae. The units of the latter type are class-dependent ones. They are acquired automatically from the iterative training procedure. At every phase of the iteration, we choose the best concatenation of morae as a new unit, according to its contribution to the drop in entropy on the training data. In the following experiments, we employed 200 of such new units per word model.

### 3.3. Implementations as N-gram Parameters

We adopted the following scheme of labeling the subword units that are used in the subword N-grams. This scheme facilitates the implementation of our hierarchical language model, requiring no change to the existing decoder [9]. The label consists of three terms: (i) the class, (ii) the starting mora position in a word, and (iii) whether or not the unit is located at the end of a word. From every unit used in the subword N-grams, we generated a set of labeled units having different labels. They were added to the lexicon and treated as pseudo-words. With these labels, we can keep the hierarchy of the original model even in the flat N-gram parameters. The first term of the label enables the subword unit to hold the different transition probabilities according to the class; the second term allows for different transitions according to the word length; and the third term allows for transitions to other classes. The third term also denotes whether a pause can be inserted after the subword.

### 4. SPEECH RECOGNITION EXPERIMENTS

To evaluate the effectiveness of the proposed modeling, we performed speech recognition experiments. The experiments compared two language models, both constructed from a primary class N-gram [8] trained on the travel arrangement corpus [6] in table 1. The two models were obtained by substituting the class N-gram at only the intra-class probability $p(w | c)$ of the Japanese Personal Names (JPN) and Japanese Family Names (JFN) classes in different ways:

- **Proposed Method:** Instead of the word 1-grams for JPN and JFN in the primary model, stochastic word models constructed from the name list [5] in table 1 are used. All the JPN and JFN words are eliminated from the lexicon as being OOV words, then a set of subword units are added to the lexicon as pseudo-words.

- **In-Vocabulary Recognition:** This shows the plausible upper bound of the proposed method. Instead of the word 1-grams for JPN and JFN of the primary model, it uses 1-grams, but from the name list [5] in table 1. All the Japanese names in the list are added to the lexicon.

The test set consisted of 550 utterances (4,990 words) in the

<table>
<thead>
<tr>
<th>Table 1: Linguistic Corpora for Model Training</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corpus size</strong></td>
</tr>
<tr>
<td><strong>Words</strong></td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
</tr>
<tr>
<td><strong>Number of words for Japanese names indicates the total number of names in the list. Vocabulary size in Japanese names is counted disregarding the difference in writings in Chinese characters.</strong></td>
</tr>
</tbody>
</table>
domain of travel arrangement [6]. It included 70 JFN/JPN words. The recognition performance was evaluated using two criteria:

- **Word Recognition Accuracy:** Whole words in utterances are evaluated. Correct recognition of JFN/JPN (i.e., OOV words in the proposed method) is defined as correct identification of readings, classes, and locations (by DP-matching), simultaneously.

- **Name Precision/Recall:** Precision and recall for only the JFN/JPN words are evaluated (based on the above DP-matching).

As shown in table 2, without any JFN/JPN words in the lexicon, the proposed method performed nearly as well as the plausible upper bound of the in-vocabulary recognition. Whole-word accuracy improved slightly as a side effect, owing to the weakness of the lexical constraints that sometimes preserves the segmentation surrounding JFN/JPN words when acoustic model mismatches can damage them.

Figure 4 shows the relationship between name (OOV) recall/precision vs. the number of automatically acquired units (other than single morae) that were adopted in the subword N-grams. The use of class-dependent subword units improved the performance almost monotonically. The 200 units at the far right correspond to the experimental conditions of table 2.

<table>
<thead>
<tr>
<th>Table 2: Speech Recognition Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recog. Rate (%)</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Word Accuracy</td>
</tr>
<tr>
<td>Name Recall</td>
</tr>
<tr>
<td>Name Precision</td>
</tr>
</tbody>
</table>

The performances under the in-vocabulary condition indicate the plausible upper bounds of the proposed method with the data set.

5. SUMMARY

We proposed a hierarchical language model incorporating class-dependent word models to cope with the problems of OOV words in continuous speech recognition. Application of the model to Japanese personal and family names showed that it performs nearly as well as the upper bound of the in-vocabulary conditions. We plan to apply the model to other proper nouns, including compound/coined/abbreviated words and hesitations in the middle of word, in the future.

6. REFERENCES