SPEECH RECOGNITION OF FOREIGN OUT-OF-VOCABULARY WORDS USING A HIERARCHICAL LANGUAGE MODEL

Hirofumi Yamamoto[1,3]*, Genichiro Kikui*[2]*, Satoshi Nakamura*[1,2]*and Yoshinori Sagisaka*[1,3]

[1]National Institute of Information and Communications Technology
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto
[2]ATR Spoken Language Communication Research Labs.
[3]GITI Waseda Univ.
{hirofumi.yamamoto, genichiro.kikui, satoshi.nakamura, yoshinori.sagisaka}@atr.jp

ABSTRACT

This paper proposes a new speech recognition scheme for foreign out-of-vocabulary words embedded in native-language speech. To recognize foreign names frequently observed in news speech or in translation speech, we adopted a hierarchical language model that had been successfully applied to OOV words covering native vocabularies. In this hierarchical language model, OOV vocabularies are modeled as a word-class model in the upper-layered model, and their statistical phonotactic constraints are modeled in the lower-layered model. Since extra statistics are needed to cover foreign words and their pronunciation differences, we have introduced two techniques. The first is to combine translation target language models and translation source statistics of OOVs using the hierarchical language model. The second is to automatically generate recognition target pronunciations from original pronunciations by syllable-to-syllable mapping. To confirm the validity of this recognition scheme, we have conducted speech recognition experiments using English speech including Japanese personal names as OOV words. The proposed method outperformed the existing algorithm using a lexicon consisting of all the words in the training set. Surprisingly, it achieved better OOV recognition results than the non-OOV condition where all the proper names in the test set are registered in the lexicon.

Index Terms: Speech Recognition, Language model, Foreign word, Out-of-Vocabulalry word, Hierarchical language model.

1. INTRODUCTION

Recently, in LVCSR, HMM acoustic models and trigram language models have become the standard practical technology. They have been widely used in many applications. In many speech technologies, speech-to-speech translation is one of the most promising applications. However, many problems in speech recognition are still to be resolved to build a practical speech-to-speech translation system. Out-of-vocabulary is one of the most serious problems for practical speech-to-speech translation systems. In human-to-human speech dialogue for speech-to-speech translation, proper nouns, such as personal names or location names, are frequently used and usually they convey important information. In the current speech recognition paradigm, all recognition targets must be registered in recognition lexicons, since only registered words can be recognition targets.

As it looks unrealistic to obtain all word entries and their statistics, we have proposed a hierarchical language model to cope with OOV words and expressions [1][2][3]. We first started to recognize OOVs consisting of single-class proper-noun OOV words and generalized to multiple OOV classes using unsupervised VQs for OOV word classes [4]. It has already been confirmed that this model can work not only for particular word classes but also OOV named entries consisting of word sequences [5]. This model enables the recognition of OOV words and expressions without registering them into recognition lexicons.

In current hierarchical language models, we have only treated the OOV words of the same language as the embedded speech. However, in speech-to-speech translation such as English-to-Japanese, not only English OOV words but also Japanese OOV words are frequently observed in English sentences. They are expected to be correctly recognized, since proper-noun OOV words often convey crucial information in speech-to-speech translation. The hierarchical language model requires OOV word statistics of frequency and their pronunciation for model training. As it is extremely difficult to collect statistics of foreign OOV words, we need some alternatives. Assignment of pronunciations to foreign OOV words is also a time-consuming and difficult task. To be free from these difficulties, we have extended the hierarchical language model as follows. In a current hierarchical language model, two types of sub-models, inter-word sub-models and intra-word sub-models, are respectively used to represent word-to-OOV-word transition probabilities and pronunciations of OOV words. The first extension is an addition to the foreign intra-word model to accept foreign languages as well as the native one. For example, a Japanese personal name inter-word model is combined with an English intra-word model. The second extension is generation of OOV word pronunciations from pronunciations of the original language by syllable-to-syllable mapping. In this mapping, multiple syllables in the target language are assigned from one syllable in the native language. These generated pronunciations have a potential of high coverage for foreign OOV word pronunciations.

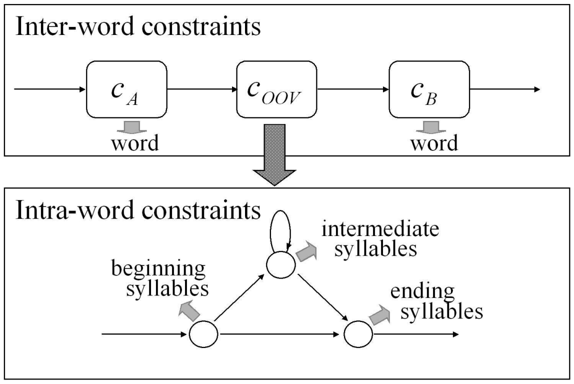

Fig. 1. A hierarchical language model
Output symbols in the upper layer and the lower layer represent word sequences and syllable sequences, respectively.

2. HIERARCHICAL LANGUAGE MODEL

In the hierarchical language model, two sub-models are used to give independent constraints to OOV words, as illustrated in Fig.1. One is an inter-word model that represents transition probabilities from words in the lexicons to and from OOV words. The other is an intra-word model that gives statistical phonotactic constraints to OOV words. These models output ”my / name / is / JH+OW+N+Z(OOV)” from input speech ”my name is Jones”, where ”Jones” is an OOV word.

2.1. An inter-word model

For the OOV inter-word model, the class-based bigrams [6] shown in the next equation are employed.

$$p(OOV|w_{i-1}) = p(c_{OOV}|w_{i-1})p(OOV|c_{OOV}) \qquad (1)$$

In this equation, OOV represents an OOV word that belongs a word class represented by c_{OOV}.

2.2. An intra-word model

For the OOV intra-word phonotactic constraint model, we employed a probabilistic FSA (Finite State Automaton) with three states. In this model, the output symbols for each state are sub-words as syllables or sub-word sequences.

Using this model, the occurrence probability of an OOV consisting of L-sub-word sequences $S^L = (s_1, s_2, ..., s_L)$ can be approximated by the following equation as a bigram:

$$p(OOV) = \begin{cases} p(s_{1,BE}) & \text{(if } L = 1\text{)} \\ p(s_{2,E}|s_{1,B}) & \text{(if } L = 2\text{)} \\ p(s_{1,B})p(s_{2,I}|s_{1,B}) & \\ \quad \prod_{i=2}^{L-1} p(s_{i,I}|s_{i-1,I})p(s_{L,E}|s_{L-1,I}) & \text{(if } L > 2\text{),} \end{cases} \qquad (2)$$

where $s_{1,B}$ and $s_{1,BE}$ respectively represent the beginning sub-words and single length sub-words. $s_{i,I}$ and $s_{i,E}$ respectively represent the i-th intermediate, and ending sub-words.

2.3. Combining inter- and intra-word models

When inter- and intra-word models are combined, $p(OOV|c_{OOV})$ in Eq.(1) is replaced by Eq.(2). After this replacement, sub-words $s_{i,[B,I,E]}$ are added to recognition lexicon as new entries. Transitions between inter- and intra-word models can be represented as from ”word to beginning sub-word” and from ”ending sub-word to word” as follows.

$$p(s_{1,B}|w) = p(OOV|w)p(s_{1,B}) \qquad (3)$$

$$p(w|s_{n,E}) = p(w|OOV) \qquad (4)$$

3. EXTENSION OF THE HIERARCHICAL LANGUAGE MODEL

3.1. Incorporation of inter-word models for foreign words

The inter-word models represent word-to-OOV-word transition probabilities. Therefore, the transition probabilities of foreign OOV words must be estimated. To avoid direct estimation of these probabilities, we use probabilities in native OOV words as in the following equation.

$$p(OOV_{foreign}|w) \approx \alpha p(OOV|w) \qquad (5)$$

Here, $p(OOV_{foreign}|w)$ represents the transition probability of the foreign OOV word from word w, $p(OOV|w)$ represents probability of native OOV, and α is a constant. This approximation means OOV word context depends only on its category of it, not on the language of its origin.

3.2. Phone mapping

The purpose of the second extension is to generate pronunciations for foreign OOV words. A target OOV's pronunciation can be presented by Eq.(2). To generate this pronunciation from the original pronunciation, we introduce mapping from the original syllables. In this mapping, multiple syllables are assigned to one original syllable, and if syllables $t_1, t_2, ..., t_n$ are assigned to original syllable s, the following equation concludes.

$$p(s) = \sum_{i=1}^{n} p(t_i, s) \qquad (6)$$

Here, $p(t_i, s)$ represents the mapping probability from original syllable s to target syllable t_i. We assume that this mapping probability is independent of syllable position. This assumption gives new a inter-word model by replacing all of s in Eq.(2) with t using Eq.(6).

4. EXPERIMENTAL SETUP

To evaluate the proposed model, we conducted speech recognition experiments based on the travel conversation task [7]. The language combination in the dialogs is English-to-Japanese, and the recognition target was Japanese proper-noun OOV words in English speech. The recognition of Japanese speech with Japanese proper-noun OOVs was also conducted to confirm the validity of the original hierarchical language model as a base line.

4.1. The inter-word model

English and Japanese inter word models were used in experiments. Models were trained using 6.1M words containing 44K lexical entries for English, and 8.7M words containing 66K lexical entries for Japanese. In both languages, we adopted Multi-Class Composite bigrams [12] with 2,000 classes and 35K word successions. These models include two OOV word classes whose categories are Japanese family and first names. The α in Eq.(5) is always set to 1/2 based on the assumption that frequencies of English and Japanese personal names are the same.

4.2. The intra-word model

Only Japanese family and first name intra-word models were used in the experiments. For the Japanese family intra-word model, 304K names and 20.4K different names were used for model training. For the first name, 295K names and 19K different names were used. The number of syllable entry was 98; since Japanese is an open-syllable language, this number is rather smaller than English. The number of syllable successions was 400. This number was selected from preliminary experiments.

4.3. Phone mapping

To generate English pronunciation of Japanese personal names from Japanese syllables, syllable mapping was manually carried out by a native English speaker. The number of Japanese syllables was 98 as stated in the previous subsection. For each Japanese syllable, one or more English syllables that are similar to the Japanese syllable are selected by a native English speaker who can speak Japanese. As a result of this mapping, two English syllables were assigned to 36 Japanese syllable entries (”k a” to ”K AA” or ”K AH”, ”f u” to ”HH UW” or ”F UW”), and four English syllables were assigned to two Japanese syllable entries (”r i” to ”R IY”, ”R IH”, ”L IY” or ”L IH”). This mapping cost is quite small, it is shorter than one hour. All $p(t_i, s)$ in Eq.(6), are assigned to be 1 to simplify the model.

4.4. Evaluation measures

Three measures are used for the evaluation. The first measure is OOV word position recall. In this measure, correct recognition of OOV words is counted if the OOV word positions in a sentence and a category are correct. We disregard pronunciation accuracy in this measure. This measure shows the accuracy of OOV word positions and categories that are the most important for some applications such as speech-to-speech translation.

The second one is OOV word pronunciation recall. In this measure, not only OOV word position and category, but also its pronunciation, is checked for correctness. The final measure is conventional word accuracy in that an OOV word is considered if its position is correctly recognized. This measure can show whether the proposed model produce negative side effects.

Table 1. Acoustic and decoding conditions

Analysis	sampling rate: 16 kHz frame length/shift: 20 ms / 10 ms feature vector:12 MFCC+ 12 ΔMFCC+Δpower
Acoustic model	HMnet by MDL-SSS [8][9][10] 3,258 states, 5 mixture components, gender-dependent models
Decoder [11]	1st pass: frame-synchronous Viterbi search 2nd pass: word lattice construction using FSA and rescoring

4.5. Experimental conditions

The other experimental conditions are shown in Table 1. In the first decoding pass, we used a hierarchical language model after converting intra-word FSA model to bigrams. A viterbi search is carried out with a simple combination of inter-word class-based bigrams and an intra-word model. In the second pass, FSA is used to prune the unreasonable transitions described above. Since the N-gram is implemented on the decoder, back-off smoothing gives non-zero probabilities of unreasonable transitions, such as a transition from a word to an intermediate sub-word.

5. EXPERIMENTAL RESULTS

5.1. Evaluation of the original hierarchical language model

As a baseline for comparison, we evaluated the original hierarchical language model. The proposed model includes three components, an English inter-word model, a Japanese intra-word model and a phoneme mapping. In the original model, only inter- and intra-word models are used. Moreover, in the original model, an English inter-word model cannot be used, since the language of the inter- and intra-word must be the same. Therefore, Japanese inter-word model were used instead of English inter-word model wherever appropriate. The evaluation set included 100 utterances including at least one Japanese personal name (a total of 110 Japanese personal names were included) in each utterance. The experimental results are shown in Table 2, the experimental condition in ”In Lexicon” is that all Japanese personal names are stored in the recognition lexicon. This condition closely corresponds to the upper limit of this modeling framework. The original hierarchical language model resulted in a slightly lower performance than ”In Lexicon” condition. Thus, we confirmed the performance of the hierarchical language model.

5.2. Evaluation of the proposed model for foreign OOVs

To evaluate the proposed extended hierarchical language model, two evaluation sets were used. The first evaluation set consists of 100 English utterances without OOV words. The purpose of this

Table 2. OOV Recognition Performance in the Original Model

	Word ACC.	Position	Pronunciation
In Lexicon	94.73	91.82 (101/110)	90.00 (99/110)
OOV Model	94.08	89.09 (98/110)	86.36 (95/110)

experiment was to observe the side effects of the proposed model. We compared word accuracy for this evaluation set with and without the proposed model. Experimental results turned out to be 92.10% for the case without proposed model and 92.27% with the proposed model. This confirms that the proposed model produces no negative side effects when no OOV words are included.

The second evaluation set consists from 79 English utterances including at least one Japanese personal name (a total of 120 names are included). The experimental conditions in ”In Lexicon” were the same as those described in the previous sub-section. An English native speaker assigned pronunciations of Japanese personal names stored in recognition lexicon. The experimental results shown in Table 3 reveal that the proposed method results in a score about three points higher for OOV word position recall and the same pronunciation recall than ”In Lexicon” condition. The error reduction rate of OOV word position recall from ”In Lexicon” condition is 19%. The reason of this improvement seems to be that the generated pronunciations have higher coverage than human-assigned pronunciations. These results show that the proposed extended hierarchical language model can give good recognition performance for foreign OOV words.

Table 3. OOV Recognition Performance in the Proposed Model

	Word ACC.	Position	Pronunciation
In Lexicon	87.87	82.50(99/120)	75.83(91/120)
Proposed Model	89.08	85.83(103/120)	75.83(91/120)

6. CONCLUSIONS

In this paper, we proposed a new recognition scheme for foreign OOV words crucially needed in speech-to-speech translation. In speech-to-speech translation, not only recognition target-language OOV words, but also foreign OOV words, improve total recognition accuracy and correctly identify information useful at the translation stage.

The proposed extended hierarchical language model for foreign OOV words only needs statistics of frequency for OOVs for the inter-word model and pronunciation of foreign OOV words for the intra-word model. This modeling can greatly decrease the data collection efforts for foreign OOVs for some specified word classes frequently observed in task corpora. Since it does not require extra effort to assign foreign word pronunciations of OOVs, it does not require the generation of any pronunciation dictionary. In the proposed extended hierarchical language model, the inter-word model for foreign OOVs is generated from a model created in the origin language. Therefore, inter- and intra-word models can be independently created for speech-to-speech translation target languages.

Furthermore, the proposed model results in higher foreign OOV recognition performance than conditions where all OOV words are stored in a recognition lexicon and pronunciations are assigned. We can confirm that the proposed model can not only greatly reduce data collection efforts but also can achieve very good OOV word recognition performance.

7. REFERENCES

[1] K. Tanigaki, H. Yamamoto and Y. Sagisaka, “A hierarchical language model incorporating class-dependent word models for OOV words recognition,” Proc. ICSLP2000, pp.123-126, 2000.

[2] S. Onishi, H. Yamamoto, Y. Sagisaka, “Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes,” Proc. Eurospeech2001, pp.693-696, 2001.

[3] Y. Ogawa, H. Yamamoto, Y. Sagisaka, G. Kikui, “Word class modeling for speech recognition with out-of-task words using a hierarchical language model,” Proc. Eurospeech2003, pp.221-225, 2003.

[4] H. Kokubo, H. Yamamoto, Y. Sagisaka and G. Kikui, “Out-of-vocabulary word recognition with a doubly markov language model,” Proc. ASRU2003, pp.543-547, 2003.

[5] Y. Tomita, Y. Okimoto, H. Yamamoto and Y. Sagisaka, “Speech recognition of a named entity,” Proc. ICASSP2005, pp.I-1057-1060, 2005.

[6] S. Bai, H. Li, B. Yuan, “Building class-based language models with contextual statistics,” Proc. ICASSP98, pp.173-176, 1998.

[7] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto and S. Yamamoto, “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world,” Proc. 3rd International Conference on Language Resources and Evaluation, pp.147-152, 2002.

[8] J. Takami and S. Sagayama, “A successive state state splitting algorithm for efficient allophone modeling,” Proc. ICASSP92 pp.573-576, 1992.

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

[10] T. Jitsuhiro, T. Matsui and S. Nakamura, “Automatic generation of non-uniform HMM topologies based on MDL criterion,” IEICE Trans. Inf. ,E87-D, pp.2121-2129, 2004.

[11] T. Shimizu, H. Yamamoto, H. Masataki, S. Matsunaga and Y. Sagisaka, “Spontaneous dialogue speech recognition using cross-word context constrained word graphs,” Proc. ICASSP96, pp.17-41, 1996.

[12] H. Yamamoto, S. Isogai and Y. Sagisaka. “Multi-class composite N-gram language model,” Speech Communication 41 (2003), pp.369-379, 2003