# Creation of Learner Corpus and its Application to Speech Recognition

**Hiroki Yamazaki, Keisuke Kitamura, Takashi Harada, Seiichi Yamamoto**
Department of Information Systems Design, Doshisha University
1-3 Miyakodani, Tatara, Kyotanabe-shi, Kyoto 610-0321, Japan
E-mail: bte7015@mail4.doshisha.ac.jp, seyamamo@mail.doshisha.ac.jp

**Abstract**

Some big languages like English are spoken by a lot of people whose mother tongues are different from. Their second languages often have not only distinct accent but also different lexical and syntactic characteristics. Speech recognition performance is severely affected when the lexical, syntactic, or semantic characteristics in the training and recognition tasks differ. Language model of a speech recognition system is usually trained with transcribed speech data or text data collected in English native countries, therefore, speech recognition performance is expected to be degraded by mismatch of lexical and syntactic characteristics between native speakers and second language speakers as well as the distinction between their accents. The aim of language model adaptation is to exploit specific, albeit limited, knowledge about the recognition task to compensate for mismatch of the lexical, syntactic, or semantic characteristics. This paper describes whether the language model adaptation is effective for compensating for the mismatch between the lexical, syntactic, or semantic characteristics of native speakers and second language speakers.

## 1. Introduction

Some big languages like English are spoken in various occasions like presentations and discussions at international conferences by a lot of people whose mother tongues are different from. Their second languages often have not only distinct accent but also different lexical and syntactic characteristics. Speech recognition performance is severely affected when the lexical, syntactic, or semantic characteristics in the training and recognition tasks differ (Bellegarda, 2004). Language model of a speech recognition system is usually trained with transcribed speech data or text data collected in English native countries, therefore, speech recognition performance is expected to be degraded by mismatch of lexical and syntactic characteristics between native speakers and second language speakers as well as the distinction between their accents. However, it is almost impossible to collect large spoken and text data of second language speakers in various discourse contexts, even though such second language speakers' corpora are indispensable for training a language model proper for their speech.

The aim of language model adaptation is to exploit specific, albeit limited, knowledge about the recognition task to compensate for mismatch of the lexical, syntactic, or semantic characteristics. Language model adaptation has been investigated in various frameworks such as a cross-domain adaptation, where a small adaptation corpus relevant to the current recognition task was created and a language model trained with them was merged into the other language model trained with a large background corpus associated with a presumably related but somewhat different task. However, it has not been explored whether the language model adaptation is effective for compensating for the mismatch between the lexical, syntactic, or semantic characteristics of native speakers and second language speakers. If the adaptation is effective for compensating such a mismatch, methodology on creating corpus of second language speakers proper in various domains should be investigated.

This paper describes some experimental results for investigating whether the language model adaptation is effective for compensating for the mismatch between the lexical, syntactic, or semantic characteristics of native speakers and second language speakers. The remainder of this paper is organized as follows. Section 2 introduces a test-bed system of English speech recognition for Japanese. Section 3 explains a learner corpus of English sentences translated by Japanese subjects whose communicative skills in English were evaluated, and also describes a language model trained with the learner corpus. Section 4 presents some experimental results that demonstrate effectiveness of the language model adaptation, and explores some relations between the performance and the methodology of creating the learner corpus, followed by concluding remarks.

## 2. A Test-bed System of English Speech Recognition for Japanese

In order to verify effect of the language model adaptation for compensating for the mismatch between the lexical, syntactic, or semantic characteristics of native speakers and those of second language speakers, the authors employed a multi-lingual speech recognition system ATRASR (Nakamura, 2006) developed at ATR as a test-bed system of English speech recognition for Japanese. Main features of ATRASR are shortly described in the following;

(1) Speech Analysis: The speech analysis conditions were as follows; the frame length was 20 ms and the frame shift was 10ms; 12-order MFCC, 12-order Delta MFCC, and Delta log power were used as feature parameters. Table 1 shows the phoneme units. Our phoneme sets consist of 44 phonemes, including silence. They are the same as those used in the WSJ corpus official evaluations because in this way we could use its dictionary as a source of pronunciation base-form.

(2) Acoustic Model: The authors employed two acoustic models; one is a native acoustic model which was trained

with speech data in Wall Street Journal (WSJ) corpus. About 37,500 utterances recommended for speaker-independent training were selected as the training set for the acoustic model (Paul, 1992). The total number of speakers is 284 (143 male and 141 female). The other acoustic model is a non-native acoustic model which was trained with read speech data collected from 201 Japanese subjects (100 male and 101 female) (Minematsu, 2003). Each subjects read out about 120 English sentences. An HMNet (Takami, 1992), a kind of context-dependent phone models, was used for modeling both acoustic models. The topology of HMNet was determined with the speech data of native speakers and the same topology was used for the English acoustic model for Japanese (the non-native acoustic model).

| |
|---|
| **AA, AE, AH, AO, AW, AX, AXR, AY, EH, ER, EY, IH, IX, IY, B, CH, D, DH, DX, F, G, HH, JH, K, L, M, N, NG, OW, OY, P, R, S, SH, T, TH, UH, UW, V, W, Y, Z, ZH,** |

Table 1: Sub-word units for English ASR.

(3) Pronunciation Dictionary: The pronunciation dictionary has about 35,000 entries mainly for vocabulary concerning travel conversations.
(4) Language Model: We have two language models; one is a native language model which was trained with BTEC (Basic Travel Expression Corpus) English data which consists of about 500,000 sentences (Takezawa, 2002). The other language model is a non-native language model which will be explained in the following section.
(5) Decoder: We used a two pass decoder ATRASR developed at ATR, which uses bi-gram language model and inter-word context dependent acoustic model in the first pass, and rescores candidates obtained in the first pass with tri-gram language model and intra-word context dependent acoustic model in the second pass.

## 3. Creation of Learner Corpus and Language Model Training

In order to verify effect of the language model adaptation for compensating for the mismatch between the lexical, syntactic, or semantic characteristics of native speakers and those of second language speakers, the authors created Japanese learner corpus of 150,000 English sentences translated by 500 Japanese subjects (Kitamura, 2007), of which Japanese source sentences were randomly selected from BTEC, English-Japanese parallel corpus of 500,000 paired sentences, and textbooks on English for junior and senior high schools. The subjects' communicative skills in English were evaluated on TOEIC (test of English for International Communication) score (toeic, 2008), the most popular English assessment of communicative English skill in Japan, which ranges from 10 (lowest) to 990 (highest). Each subject was requested to submit his or her score record certification of TOEIC before translating Japanese sentences to English. The subjects were selected so that their TOEIC scores were averagely distributed from 300 to 990. The subjects were divided into five groups so that their TOEIC scores in each group were distributed equally. 300 Japanese sentences were shown by sentence-by-sentence to each subject via a terminal display, and the subject was requested to translate them to English without consulting a dictionary. Therefore, each Japanese source sentence has 100 translations by different subjects. After finishing translating all source sentences, only misspellings in the translated sentences were checked and corrected.

The total sentences of the learner corpus are 150,000, while the BTEC has about 500,000 English sentences. The vocabulary of the learner corpus is 8,441, while that of the BTEC corpus is about 25,000. Various abusage like lack of indefinite articles and mismatch of tense are found in many translated sentences in the learner corpus.

We used Katz back-off smoothing technique to create bi-gram and tri-gram language models by using the learner corpus, and linearly interpolated them with original language models which were trained with the BTEC English corpus. Weighting factors of both language models were determined with a cross-validation method. We call the interpolated language model a non-native language model.

## 4. Speech Recognition Experiments

We made an English test set (Nakai, 2007) which was the collection of utterances of other 54 Japanese subjects (27 female and 27 male) who were asked to translate orally 50 Japanese sentences selected from the BTEC corpus and the English textbooks. We call the test set "spontaneous speech test set" (SSTS). After finishing translating all source sentences, only misspellings in the translated sentences were checked and corrected. The subjects were also requested to read 180 English sentences selected randomly from the BTEC. We call the speech set "read speech test set" (RSTS). Figure 1 shows a distribution of TOEIC score of the subjects.

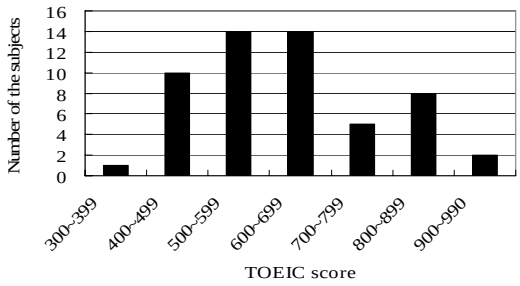


Figure 1: Distribution of TOEIC score of subjects.

As already mentioned before, two acoustic models were trained for the speech recognition system. One is an acoustic model trained with 37,500 utterances of 284 English native speakers, and the other is an acoustic model trained with English read speech collected from 201 Japanese subjects. Table 2 shows the word accuracy for the read speech test set with the native acoustic model and with simultaneous use of both acoustic models. As a reference word accuracy averaging over speech by two English natives is also shown in the Table. The

simultaneous use of both acoustic models improved word accuracy for the read speech test set (RSTS) by 43 percent for male voices and by 36 percent for female voices at average in comparison with the acoustic model trained with English native speakers.

| | native acoustic model | both models |
|---|---|---|
| male | 29.8% | 73.6% |
| female | 36.7% | 72.5% |
| natives | 82.3% | 53.5% |

Table 2: Word accuracy of RSTS with the native acoustic model and with both acoustic models.

In the following experiments, we used both acoustic models simultaneously because the best word accuracy of averaging over the subjects was obtained for the case that both acoustic models were used simultaneously. The simultaneous use of both acoustic models improved word accuracy for the read speech test set in comparison with the acoustic model trained with English native speakers. However, the word accuracy for orally translated speech test set (SSTS) is at average about 10 percent lower than that read speech test set (RSTS). There are still a lot of room for improving performances of both acoustic models and language models.

### 4.1 Language Model Adaptation

First we tried to use language adaptation technologies. Table 3 shows comparisons between the test set perplexity of the spontaneous speech test set (SSTS) with the interpolated language model (non-native language model) and the original native language model.

| | native language model | non-native language model |
|---|---|---|
| bi-gram | 50.7 | 34.9 |
| tri-gram | 23.9 | 13.9 |

Table 3: Perplexities of the native language model and the non-native language model for SSTS.

The reason why the test set perplexity decreases could be explained with two reasons; one is the interpolated language model is better representation of English utterances by Japanese, and the other is effect with increase of language data. In order to assure the reason, we tried to measure the test set perplexity of the read speech test set (RSTS). Table 4 shows the experimental results. As is clearly shown in the table, the interpolated language model does not decrease the perplexity of RSTS. Therefore, the decrease of the perplexity is due to better representation of SSTS, not to increase of language data.

| | native language model | non-native language model |
|---|---|---|
| bi-gram | 43.9 | 49.1 |
| tri-gram | 22.7 | 23.6 |

Table 4: Perplexities of the native language model and the non-native language model of RSTS.

Figure 1 shows the relation between TOEIC score of each subject and decrease of perplexity measured with bi-gram for his or her utterances. Decrease of perplexity by the interpolated language model is more for the utterances spoken by the subject of lower TOEIC score than the utterances by the subjects of higher TOEIC score, as is shown in Fig. 1, though the value widely changes depending on each subject.

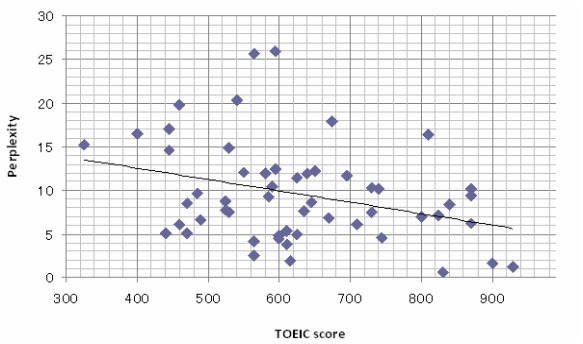


Figure 2: Relation between TOEIC score of each subject and decrease of perplexity for his/her utterances.

### 4.2 Word Accuracy

Speech recognition experiments were carried out under conditions that various parameters concerning the decoder such as beam width were kept the same. Table 5 shows the comparisons between word accuracies for SSTS with the native language model and the non-native language model.

| input sentences | native language model | non-native language model |
|---|---|---|
| BTEC | 69.4% | 73.7% |
| textbooks | 60.1% | 78.9% |
| SSTS | 62.5% | 77.5% |

Table 5: Comparisons between word accuracies for SSTS with the native language model and the non-native language model.

The non-native language models (bi-gram in the first pass and tri-gram in the second pass) improved the word accuracy by about 15 percent in comparison with the native language models. The improvement by the interpolated language model is less for speech by the subjects of higher TOEIC score than speech by those of lower score, as is shown in Fig. 3. As is clearly shown in Fig. 3, the non-native language model improved word accuracy by more than 10 percent even for speech of the

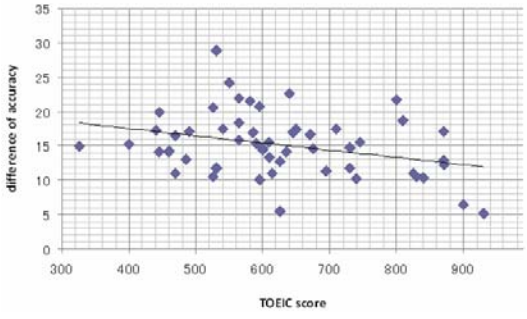


Figure 3: Relation between TOEIC score of each subject and improvement of accuracy for his/her utterances.

subjects of higher TOEIC score who usually use longer expressions than the subjects of lower TOEIC score.

### 4.3 Methodology of Corpus Creation

As is clearly shown in the previous discussions, the language model which was trained with the learner corpus, decreases perplexity of utterances in the same domain and increases the recognition accuracy. This result shows that the language model adaptation is effective to compensate for lexical and syntactic characteristics of second language speakers as well as the acoustic model adaptation is useful for compensating for distinct accent. As Lefevre et al. demonstrated (Lefevre, 2001), out-of-domain speech training data do not cause significant degradation of the system performance. On the contrary it was found to be more sensitive to the language model domain mismatch. As mentioned before, however, it is almost impossible to create a learner corpus in various discourse contexts.

We have tried to investigate relations between quantity and quality of the learner corpus and improvement of performance. Figure 4 shows the relation between perplexity of SSTS and quantity and quality of the corpus. The vertical axis represents perplexity with bi-gram, and the horizontal one shows each case in which quantity of the corpus changes. The black bars depict cases that number of task sets changes and the white bars represent cases that number of subjects changes.

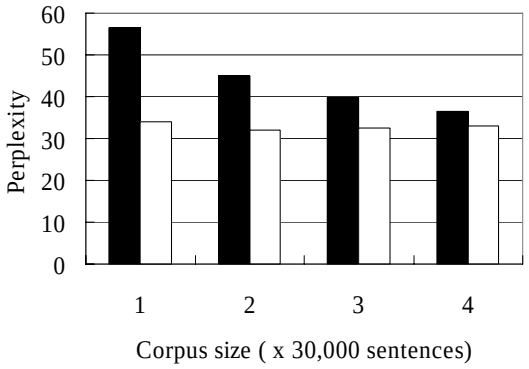


Figure 4: Perplexity with bi-gram for each case that quantity and quality of the corpus change.

As is clearly shown in Fig. 4, language models with learner corpus of more target sentences leads to better performance in comparison with that of more subjects. Therefore, we may be able to reduce the number of subjects, if we can select adequate candidates of the target sentences.

## 5. Conclusions

Speech recognition performance was improved with the interpolated language model between the original language model and the language model trained with the learner corpus which was collected in the same domain as the BTEC corpus. This result shows that the language model adaptation is effective to compensate for lexical and syntactic characteristics of second language speakers as well as the acoustic model adaptation for compensating for distinct accent. As is already described, some big languages like English are spoken in various occasions like presentations and discussions at international conferences by a lot of people whose mother tongues are different from. Such speech will be recorded, transcribed with speech recognition systems, and re-used. However, their second languages often have not only distinct accent but also different lexical and syntactic characteristics, and may be miss-recognized with the speech recognition systems which were trained with transcribed speech data or text data collected in English native countries.

As Lefevre et al. demonstrated, out-of-domain speech training data do not cause significant degradation of the system performance. On the contrary it was found to be more sensitive to the language model domain mismatch. As mentioned before, however, it is almost impossible to create a learner corpus in various discourse contexts. Various abusage like lack of indefinite articles and mismatch of tense found in the learner corpus are expected to be common in various discourse contexts, and the learner corpus may be used for language model adaptation in various domains. We plan to examine portability of the language models trained with the learner corpus, by interpolating them with a language model trained with the corpus collected in other domain.

## 6. Acknowlegements

The research reported here was supported in part by a contract with MEXT number 16300048. The authors thank to Professor M. Yanagida, of Doshisha University and Dr. K. Yasuda for various discussions.

## 7. References

Bellegarda, J.R. (2004). *Statistical language model adaptation: review and perspectives*, Speech Communication, Vol. 42 No. 1, pp. 93-108.

Kitamura, K, Yasuda, K. Yamamoto, S., and Yanagida M. (2007). *Development of learners' Japanese-to-English translation corpus*, IEICE Conf. D-15-25, (in Japanese).

Lefevre, F., Gauvain, J.L., and Lamet, L. (2001). *Improving genericity for task-independent speech recognition*, in Proc. Eurospeech2001, pp.1241-1244.

Minematsu, N., Nishina, K., and Nakagawa, S. (2003). *Readspeech database for foreign language learning*, JASJ, Vol. 59, No. 6, pp. 345-350 (in Japanese).

Nakai, Y., Yasauda, K., Yamamoto, S., and Yanagida, M.(2007). *Evaluation of Japanese English skill with English acoustic model*, IEICE Conf. D-15-31, (in Japanese).

Nakamura, S., et al (2006). *The ATR Multilingual Speech-to-Speech Translation System*, IEEE Trans. ASLP, Vol. 14, No.2 pp. 365-376.

Paul, D., and Baker. J. (1993). *The design for the wall street journal-based CSR corpus*, in Proc. DARPA Speech and Natural Language Workshop, pp. 357-362.

Takami, J., and Sagayama, S. (1992). *A successive state splitting algorithm for efficient allophone modeling*, in Proc. ICASSP, Vol. 1, pp. 573-576.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). *Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world*, in Proc. LREC, 27-2, pp.147-152.

TOEIC (2008). http://www.ets.org/toeic/.