NON-NATIVE ENGLISH SPEECH RECOGNITION USING BILINGUAL ENGLISH LEXICON AND ACOUSTIC MODELS

S. Matsunaga, A. Ogawa, Y. Yamaguchi, A. Imamura

NTT Cyber Space Labs.,
1-1, Hikarinooka, Yokosuka, Kanagawa 239-0847 JAPAN

ABSTRACT

This paper proposes an English speech recognition system which can recognize both non-native (i.e. Japanese) and native English speakers' pronunciation of English speech. The system uses a bilingual pronunciation lexicon in which each word has both English and Japanese phoneme transcriptions. The Japanese transcription is constructed considering typical Japanese pronunciation of English. Japanese and English acoustic models are used in recognizing both transcriptions, and the highest-likelihood word sequence obtained in combining with native English- and Japanese-pronounced words is the recognition result. Continuous speech recognition experiments show that the proposed system greatly improves Japanese-English speech recognition performance while maintaining the same performance level as that of a purely native English recognition system.

1. INTRODUCTION

Speech recognition of Japanese English (i.e. English spoken by Japanese people with Japanese accents) is particularly useful for improving computer-assisted learning programs that detect mispronunciation. To achieve a better human-machine interface, information retrieval systems that employ speech such as Voice Portal, require speech recognition capacity that will enables them to flexibly recognize English spoken by both native and non-native speakers.

There has been many studies in non-native speech recognition using acoustic adaptation or lexical modeling [1-5]. Although most Japanese speak with unique Japanese pronunciation, there are many Japanese who speak the language as well as native speakers do. In other words, the variety of Japanese English speech is quite wide, and so even if a recognition system uses such a lexicon employing Japanese English pronunciation dictionary, or acoustic models generated using Japanese English utterances, it can accurately recognize the speech of a limited number of people, usually at a sacrifice of its ability to recognize the speech of native speakers and Japanese who are fluent in English.

To address this problem, we propose a Japanese and native English speech recognition system that uses a bilingual pronunciation lexicon in which each word has both English and Japanese phoneme transcriptions. For example, the word "probably" is written as */p r a b ax b l iy/* in English transcription, and */p u r o b a b u r i i/* in Japanese. Each phoneme symbol is recognized both Japanese or native speaker acoustic models, and the highest-likelihood word sequence, which contains a combination of English-pronounced and Japanese-pronounced words, is derived as recognition results. We incorporated the method into the system that assesses a penalty on inter-lingual pronunciation transition, which prevents frequent transition between Japanese and English pronunciation. We also designed a Japanese English speech database to evaluate the system, and examined the applicability of the lexicons and acoustic models by comparing the results of speech recognition accuracy for Japanese and native English utterances. Experiments were also carried out using speech data containing both Japanese and native-speaker utterances.

2. DATABASE OF JAPANESE ENGLISH SPEECH

We prepared the databases described below to evaluate the performance of speech recognition for Japanese English and English spoken by native speakers.

[Data 1] English sentences uttered by the Japanese
For speech contents, we selected phonetically-balanced 1,000 sentences from LDC North American Corpus and divided them into 10 sets. There were 100 male and 100 female subjects, each of whom uttered 100 sentences. All of the subjects were confident in their ability to speak English, and their age ranged from late teens to early 50s (average age was 25.9 years).

[Data 2] English sentences uttered by native speaker
Using the same contents as Data 1, 50 male and 50 female native speakers of North American English participated in this recording. Each set of 100 sentences was uttered by 5 males and 5 females. Their age ranged from late teens to early 60s (average age was 30.1 years).

3. RECOGNITION METHODS USING JAPANESE AND ENGLISH MODELS

We used five different recognition methods to compare each level of recognition accuracy. Methods 1 and 2 used a monolingual pronunciation lexicon. Methods 3, 4 and 5 used a bilingual pronunciation lexicon. Our proposed methods are Methods 4 and 5. In this experiment, the Japanese and English phoneme models are different models. (i.e. English /*p*/ and Japanese /*p*/ are considered to be different models). However, we used the same silence models for both. A block diagram of each method is shown in Figure 1.

[Method 1:M1]

We used the lexicon with an English phoneme transcription only (i.e. "probably"=/*p r aa b a b l iy*/).

[Method 2:M2]

We used the lexicon with a Japanese phoneme transcription only (i.e. "probably"= /*p u r o b a b u r i i*/). The Japanese phoneme transcription is constructed using typical Japanese accents in speaking English. Corresponding to the diversity of utterances of Japanese people, the number of words with multiple pronunciations was greater in the Japanese than in the English phoneme transcription.

[Method 3:M3]

This method is a combination of Methods 1 and 2. The results of speech recognition using only English phoneme transcription and only Japanese transcription were compared, and the highest-scoring result was denoted as the final result. The processing amount is that of Method 1 plus Method 2, and so this is the most time-consuming method and impractical for real time systems.

[Method 4:M4]

We constructed a new lexicon by integrating the lexicons used in Methods 1 and 2, and so each word entry has both English and Japanese phoneme transcriptions. By using this lexicon, the recognizer can process word-by-word using native acoustic models or Japanese models, and chooses the most accurate phoneme transcription as a result.

[Method 5:M5]

This method also uses a bilingual lexicon. A penalty was assessed each time a phoneme chosen by the recognizer transited from one language to another. As in Method 4, the phoneme transcription with the highest score is selected as a result, but in this method we were able to avoid frequent inter-language word transitions by assessing penalties. In this experiment, back-off probabilities [6] were employed as a penalty for inter-language word transitions instead of word bigram or trigram probabilities.

The recognition system used in these methods employed a two-pass search strategy[7]. In the first pass, a time synchronous beam search using word bigrams, native-speaker (and/or Japanese) intra-word triphone HMMs, and native-speaker (and/or Japanese) inter-word monophone HMMs is carried out to generate a word lattice. In the second pass, a more accurate search using word trigrams and native-speaker (and/or Japanese) intra- and inter-word triphone HMMs is carried out to derive recognition results from the word lattice. Inter-lingual triphone HMMs (i.e. *u*(J)–*p*(E)+*r*(E); (E) is an English (native-speaker) phoneme and (J) is a Japanese phoneme.) are not used in Methods 4 and 5.

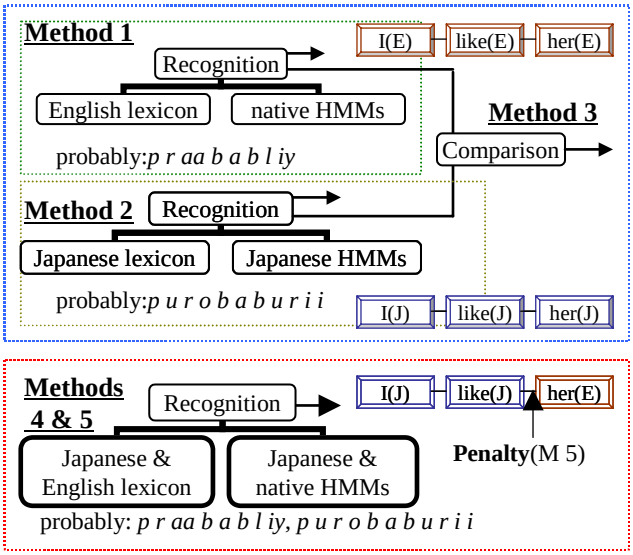


Figure 1. Block diagram for each recognition method

4.EVALUATION EXPERIMENTS

4.1. Effectiveness of Japanese phoneme transcription and Japanese phoneme model

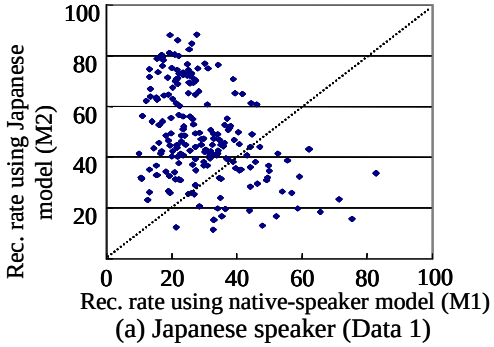
Native-speaker and Japanese phoneme HMMs, which are 3-state 8-mixture, context-dependent, speaker-independent acoustic models, are generated as baseline models. The native-speaker phoneme model has 43 phonemes and was trained with 49k utterances from LDC Resource Management and the Wall Street Journal. The Japanese phoneme model has 31 phonemes and was trained with phonetically balanced Japanese words, and sentences and loan words uttered by 400 speakers.

We selected 70-80 from among 100 sentences in each data set to make the perplexity almost equal. We evaluated these sentences as uttered by 200 speakers from Data 1. The amount of vocabulary in the experiments was 5k. In language modeling, the training data contained the same sentences as those in the evaluation data because the language model was trained using 1800k sentences from LDC North American News Corpus. We also evaluated Data 2 to compare between native and Japanese English.

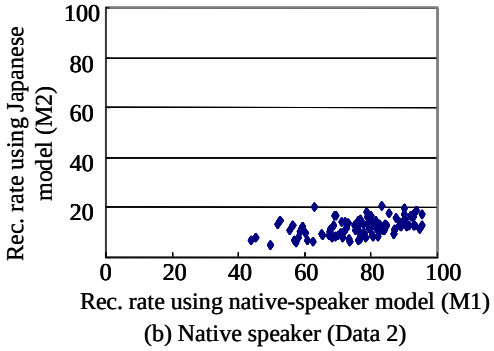
According to the evaluation results (Table 1), it is effective to apply the Japanese phoneme transcription and the Japanese phoneme models to the speech recognition of Japanese English (M1 < M2). The recognition rate was more accurate when we used both Japanese and English phoneme models (M2 < M3, M4, M5). On the other hand, applying the Japanese phoneme models to the speech recognition of native speakers tends to lower the recognition rate (M1 > M3, M4, M5). Of all the methods used, Method 4 lowered the accuracy rate most. We consider that this is because there is no restriction on inter-language word transition in this model.

Table 1. Recognition performance for each method using baseline acoustic models (%correct)

Data \ Method	M1(E)	M2(J)	M3	M4	M5
Data-1(J)	29.1	49.2	52.7	52.8	53.9
Data-2(E)	76.2	12.1	73.9	61.8	70.5



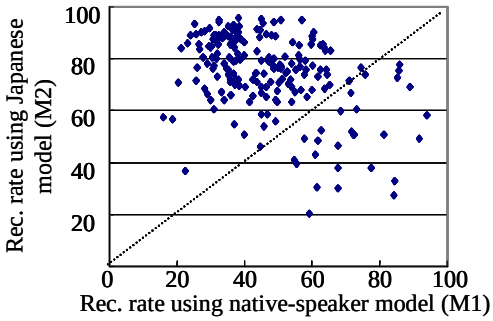
(a) Japanese speaker (Data 1)



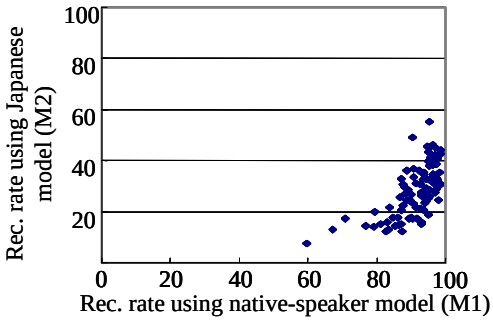
(b) Native speaker (Data 2)

Figure 2. Correlation of performance using native-speaker vs. Japanese baseline models

Figure 1 compares recognition rates, with the y-axis showing the rate for 200 speakers using the lexicon with Japanese phoneme transcription, and the x-axis showing that using the lexicon with English phoneme transcription. Each dot indicates an individual speaker. The results showed a significant difference between (a) Japanese speakers and (b) native speakers. It is clear that many Japanese speakers were influenced by the way Japanese is pronounced when they speak in English, and that the variety in their English pronunciation is consequently much greater than that of native speakers. About one-fifth of the speakers achieved higher accuracy by using native-speaker models.



(a) Japanese speaker (Data 1)



(b) Native speaker (Data 2)

Figure 3. Correlation of performance using native-speaker vs. Japanese adapted models

4.2. Task adaptation of Acoustic Model

4.2.1. Experiments using adapted models

Next, MAP adaptation [8] of acoustic models was carried out to examine the effect of the proposed methods. In doing so, we used more adequate and elaborate acoustic models than those used in the experiments described in Section 4.1. We adapted the English phoneme models using native utterances from Data 2 and the Japanese phoneme models using Japanese English utterances from Data 1. The adaptation data included neither the same transcripts nor the same speakers as those used in the evaluation data. We conducted the speech recognition experiment using both the adapted native-speaker and Japanese models.

The results, shown in Table 2 and Figure 3, show that the recognition rate of Japanese English improved significantly (from 49.2% to 74.2%) when we applied Japanese phoneme models that were adapted using English speech uttered by Japanese speakers, and that the recognition rate obtained using both Japanese and native-speaker phoneme models was greater than that when only Japanese models were used (M2 < M3, M4, M5). We also proved that Method 5 is suitable for both Japanese and native English speaker since it maintains the accuracy rate of speech recognition for native speakers (i.e. 91.4%[M1]

vs. 92.8%[M5]). This result is contrary to that shown in Table 1 and shows that Method 5, which uses constraints on inter-lingual word transition, is useful when adequate phoneme models are used.

Table 2. Recognition performance using adapted acoustic models (%correct)

Data \ Method	M1(E)	M2(J)	M3	M4	M5
Data-1(J)	46.8	74.2	78.1	76.9	80.0
Data-2(E)	91.4	28.2	91.3	85.6	92.8

Table 3. Subjective opinion criteria

Value(r)	Criterion
1	Speaking with Japanese vowels and consonants
2	Sounds a little like Japanese
3	Each phoneme is distinguishable but not as clear as native speaker
4	Good pronunciation by a English learner
5	Fluent English; Sounds like non-Japanese

Table 4. Spoken English ability performance (%correct)

Subject \ Method	M1(E)	M2(J)	M3	M4	M5
r<=2 (13%)	34.0	82.9	83.1	80.1	85.8
2<r<=3 (45%)	39.3	79.9	80.2	78.5	82.8
3<=r<4 (30%)	54.6	69.4	73.8	74.2	75.4
4<r (12%)	71.1	54.4	75.3	73.1	73.9

4.2.2. Performance comparison corresponding to English skill

Next, we carried out a subjective opinion test using two bilingual persons. In this test, the Japanese speakers of Data 1 were evaluated from rank (r) 1 to 5 based on their English skill. Skill criteria are listed in Table 3. In this table, for example, speakers of rank 5 can speak like most native-like speakers, and so all speakers are divided to four groups using the averaged subjective opinion value.

The recognition rate for each skill is listed in Table 4. This table shows that there is a strong correlation between subjective opinion and recognition rate. More Japanese-style speakers achieve higher recognition rate using Japanese models, and more native-like speakers do the same by using native models. It is quite clearly shown that of all the skill groups, Method 5 achieves a higher rate than that of Method 2 that uses only Japanese models, and than Method 1 that uses only native models.

4.3. Evaluation of native-speaker and Japanese mixed speech

Finally, we composed speech data (Data 3) that contains of both native-speaker and Japanese speech utterances. Each unit of speech data is a sequence of a native-speaker sentence utterance of Data 2 and a Japanese sentence utterance of Data 1. In this case, we prepared about 7,500 speech samples uttered by 100 native speakers and 100 Japanese speakers. Recognition performance is given in Table 5. The results show that the proposed methods achieved higher accuracy, and that Method 5 achieved slightly higher accuracy than Method 4, showing the usefulness of transition constraints.

Table 5. Recognition rate for native-speaker and Japanese mixed data (%correct)

Data \ Method	M1(E)	M2(J)	M3	M4	M5
Data-3(E+J)	49.7	48.9	49.6	78.9	81.1

5. CONCLUSION

This paper proposes a Japanese English speech recognition method using Japanese and English phoneme transcription lexicons, and native-speaker and Japanese acoustic models. This method uses transition constraints between inter-language word sequences. Experimental results showed that the proposed system can achieve higher accuracy than a system using a Japanese lexicon and acoustic models to recognize Japanese English. The results also show that, for native speakers, the proposed method has almost same ability to recognize spoken English as a system which uses an English lexicon and native-speaker acoustic models when acoustic models are adequately trained

In future work, we plan to examine the most suitable penalty that should be assessed in cases of inter-lingual pronunciation, and how effective non-native speaker adaptation would be carried out.

6. REFERENCES

[1] S. Goronzy, et al., “Is non-native pronunciation modeling necessary? ,” Proc. Eurospeech-01, pp.305-308, 2001

[2] S. Witt and S. Young, “Off-line acoustic modeling of non-native accents,” Proc. EuroSpeech’99, pp.1367-1370, 1999

[3] X. He, and Y. Zao, “Fast model adaptation and complexity selection for non-native English speakers,” Proc. ICASSP-02, pp.I-577-580, 2002

[4] V. Fisher, E. Janke, and S. Kunzmann, “Likelihood combination and recognition output voting for the decoding of non-native speech with multilingual HMMs,” Proc. ICSLP-02, pp.489-492,, 2002.

[5] N. Minematsu, et al., “Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition,” Proc. ICSLP-02, pp.529-531,, 2002.

[6] S. M. Katz, “Estimation of probabilities from sparse data for language model component of a speech recognizer”, IEEE Trans. ASSP, vol.35, pp.400-401, 1987

[7] A. Ogawa, Y. Noda, and S. Matsunaga, “Novel two-pass search strategy using time-asynchrnous shortest-first second-pass beam search,” Proc. ICSLP-00, 2000.

[8] J. Gauvain and C.H. Lee, “Speaker adaptation based on MAP estimation of HMM parameters,” Proc. ICASSP-93, pp.558-561, 1993