A COMPARATIVE ANALYSIS OF JAPANESE AND ENGLISH DIGIT RECOGNITION

Kazuhiro Kondo[*], *Joseph Picone*[**], *and Barbara Wheatley* [**]

Texas Instruments

[*]Tsukuba Research and Development Center, 17 Miyukigaoka, Tsukuba, Ibaraki 305, Japan
[**]Central Research Laboratories, P.O. Box 655474, MS 238, Dallas, Texas 75265

ABSTRACT

This paper presents initial results of comparisons between fluently spoken Japanese and English on a common task: speaker independent digit recognition with applications in voice dialing. The complexity of this task across these languages is comparable in terms of lexicon size and perplexity of the language model. The English lexicon contained 11 words, and the Japanese lexicon contained 13 words. The durations of the words, as well as phones proved to be longer and have greater variation in English than in Japanese. An analysis of several key recognition parameters, namely the frame duration, LPC order, and feature vector dimensionality are also included. None of the above parameters seems to show language dependency in our test.

1. INTRODUCTION

With increasing needs and opportunities for speech technology applications in a variety of languages, there is a growing interest in systems capable of recognizing multiple languages. However, it is obvious that some language dependencies exist in the recognition of speech. In [1], the language differences between French and English are explored on a large vocabulary recognition task. We will attempt a comparison of Japanese and English recognition with a more specific task: digit recognition.We will try to analyze the language dependent aspects of this task through experimentation. In this manuscript, recognition performance of both languages using word models and phone models will be presented. Sensitivity analysis of several key recognition parameters, with an emphasis on acoustic processing parameters will be also given.

2. COMPARISON OF THE CORPORA

The work reported here is based on two speech corpora: a Japanese corpus collected in Japan, at Texas Instruments facilities in Miho and Tokyo [2], and a heterogeneous American English corpus. The Japanese speech is office-quality, collected with a table-mounted linear microphone and a low to moderate level of ambient noise. The English corpus consists of two corpora collected separately. The first was collected in the United States, at facilities in Dallas, Texas (we will call this the English-Voice Dialer corpus, or English-VD for short). This was collected through a telephone handset, and the ambient conditions were comparable to the Japanese corpus. The remaining and the dominant portion of the English corpus were data from the Voice Across America (VAA) corpus [3], which was collected over public telephone lines and thus included considerably more channel noise, as well as background noise. Both the Japanese and English-VD corpora contain read speech relevant to voice dialing applications. The sentence types consisted of digits and some command phrases (such as "call home"). We used all types of sentences for training, but used only the digit sequences for our comparative analysis. The VAA corpus contains read digits and TIMIT sentences. Only data which included pure digits, and did not include significant amount of channel noise was selected for training as well as testing.

The Japanese language corpus consists of 221 speakers, 112 men and 109 women. Each speaker spoke 100 sentences, of which 1/2 were digit sentences. The English-VD corpus consists of a total of 208 speakers, 107 men and 101 women. Each speaker spoke 50 sentences, of which about 1/4 contained digit sequences. The VAA database consists of 464 male and 720 female speakers.

Table 1 summarizes the content of the corpora. Each digit in the English corpus appears about 10,000 times, while each Japanese digit appears about 14,000 times. Contrary to our expectations, the multiple readings of the digits, "shi" for 4, "shichi" for 7, and "ku" for 9 did not appear frequently enough in the corpora (even though we did not restrict our speakers from pronouncing them as such). These would have made the recognition task much harder since they are all highly confusable with other digits. However, we encountered instances where the last vowel was not articulated ("ich" for 1, "rok" for 6, and "hach" for 8). These need not be distinguished as different words, but require to be modeled separately. Thus, the required number of models for representing 13 Japanese words was 16, while 11 models were used for 11 English words.

3. CROSS-LANGUAGE COMPARISONS

3.1. Description of the Recognition System

The recognition system used here is an LPC-based

Table 1: Comparison of Words in the Corpus

Digit	English vocabulary		Japanese vocabulary	
	possible words	occurrence	possible words	occurrences
1	one	10231	ichi	11247
			ich[1]	2568
2	two	10609	ni	13743
3	three	10227	san	13745
4	four	10153	yon	13843
			shi	0
5	five	10023	go	13798
6	six	10199	roku	11768
			rok[a]	2123
7	seven	10184	nana	13891
			shichi	0
8	eight	9924	hachi	11067
			hach	2702
9	nine	9620	kyuu	13797
			ku	5
0	zero	5838	zero	12248
	oh	5835	maru	1215
			rei	353
delimiter	-	-	no	2190

[1] last vowel not articulated

HMM recognizer [4]. Speech is sampled at 8 kHz, LPC analysis is applied, and the LPC parameters are transformed into a feature vector. The feature vector is composed of spectral energy vectors output from a filter bank consisting of 14 mel-spaced filters, the short-term differences of these spectral energies, the speech level, and some voicing indicators. The total number of elements is 34. A linear transformation designed to normalize the covariance statistics of the feature vector is applied, and the least significant number of features are dropped. A unimodal Gaussian continuous distribution model is used along with a Viterbi-style maximum likelihood path scoring in the HMM model.

Both word and context-independent phone models were used in the following tests. For word models, 80% of the speakers were designated as training sets. Of the remaining 20% of the speakers, the digit sentences were used for the recognition tests (2068 Japanese sentences and 1267 English sentences). English phone models were trained on read TIMIT sentence in the VAA corpus. A total of 2766 female and 2456 male utterances were used. The Japanese phone models were trained on Acoustical Society of Japan (ASJ) continuous speech corpus [5]. We used the read ATR 503 sentences from this corpus. During training, 3317 female and 3518 male utterances were used. For the Japanese digit recognition task, 18 phones out of 28 phones was used; for English, 22 out of 46 phones were used.

3.2. Comparison of the Durational Behavior

Fig. 1 shows the distribution of word duration in each language for all the words in the vocabulary. These statistics were calculated from time alignment of each word in supervised recognition with word models. A tenth-order LPC analysis was used here. The dimension of features used was 20, the frame duration 20 msec, and a 30 msec Hamming window was applied.

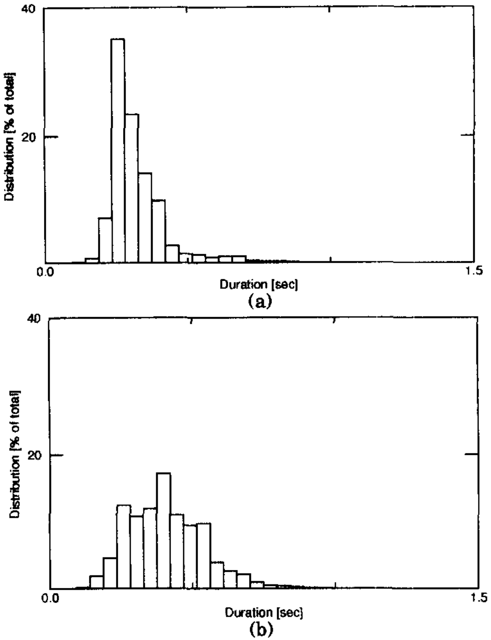
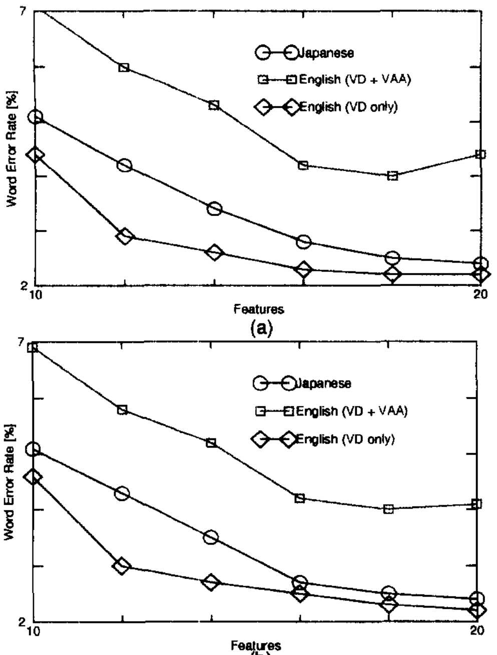


Fig. 1. Distribution of Word Duration in (a) Japanese and (b) English

Fig. 1 shows that the duration of words in the English vocabulary is larger and deviate more than its Japanese counterpart. In fact, experiments with phone models also showed that the phone durations in both language also show similarly that Japanese phones tend to be shorter and less variable. These results seem to suggest that English would generally have more variations in the path the alignment will take through each model.

3.3. Acoustic Processing Parameters

In this section, we analyze the effect of various acoustic processing parameters on the recognition performance. Since the Japanese corpus was collected with a microphone, and the English corpus was collected with telephone handsets, and since the ambient conditions differ, the absolute performance itself cannot be directly compared. Thus, we will focus our attention on the changes that occur in the performance when each acoustic processing parameters are changed. For English, recognition performance for both the entire heterogeneous corpus (English-VD + VAA) and only the



(a)

(b)

Fig. 2. The Effect of Number of Features on the Recognition Performance with (a) 10-th Order LPC Analysis and 14-th Order LPC Analysis

English-VD corpus will be presented. The former will be a more noisier corpus compared to the Japanese corpus, and the latter will be comparable or better in terms of SNR, but will be a considerably smaller corpus.

3.3.1. Feature Dimension

The dimensions of the feature vector were varied to test its effect on the recognition error rate. Fig. 2 shows the comparisons. For the Japanese and English heterogeneous corpus (English-VD+VAA), the recognition errors decrease monotonically until approximately 16 features. The slight improvement in performance beyond 16 features are not statistically significant and do not justify the increased complexity.

For the English-VD subset, however, the error rate does not decrease substantially for feature dimensions larger than ten. The recognition error rate for English in this corpus is generally similar to the error rate for Japanese, especially for dimensions over 16. Thus, for the case of comparable SNRs, comparable recognition performance is achieved. It is also worth mentioning that the performance curve was remarkably similar regardless of the LPC analysis order. This will be investigated in more detail in the next section.

3.3.2. LPC Order

Fig. 3. illustrates the effect of the LPC analysis order on the recognition performance. The feature dimension for the results shown here was 16. However, the basic trend was similar for all dimensions tested (10 to 20). For both languages, word error rate does not seem to differ significantly for any LPC analysis orders tested here (8 to 14). However, the recognition performance for the heterogeneous English corpus (English-VD + VAA) seems to degrade at an LPC analysis order of 8. This may be caused by the amount of noise included in the data, which is out of scope of this paper.

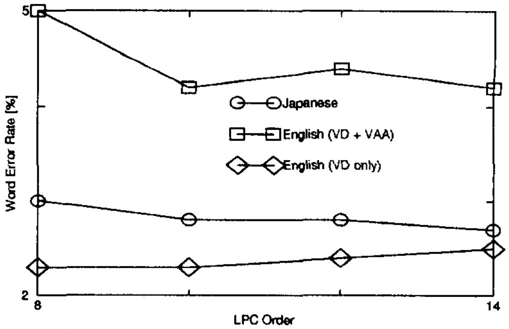


Fig. 3. The Effect of LPC Analysis Order on the Recognition Performance

3.3.2. Frame Duration

The sensitivity of the recognition error rate to frame duration was also investigated. The results are shown in Fig. 4.

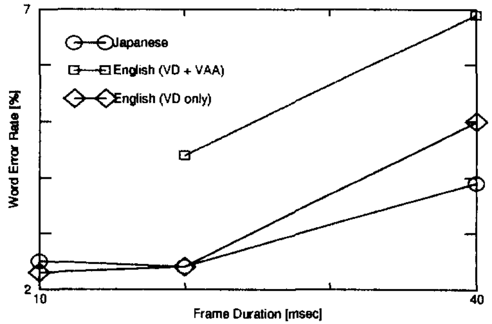


Fig. 4. The Effect of Frame Duration on Recognition Performance

Generally for both languages, the error rate did not differ for a frame duration under 20 msec. Errors increased considerably with longer frame durations. The reason for this is obviously the loss of temporal resolution (undersampling of the changing acoustics).

3.4. Error Analysis

Tables 2 and 3 show the three most common errors using word models and phone models respectively. With word models in all categories, the short tokens dominate the errors. This is understandable since these short

Table 2: Common Errors (word models)

Order	Substitutions				Deletions				Insertions			
	Japanese		English		Japanese		English		Japanese		English	
	Error	Percent[1]	Error	Percent	Error	Percent	Error	Percent	Error	Percent	Error	Percent
1	go->no	12.5	four->oh	8.1	yon	4.1	eight	6.3	ichi	9.4	oh	6.5
2	yon->no	4.1	three->two	4.4	nana	3.9	oh	4.8	rei	3.3	nine	4.8
3	zero->rei	2.5	two->zero	3.4	ni	3.9	six	2.4	roku	2.5	two	3.4

Table 3: Common Errors (phone models)

Order	Substitutions				Deletions				Insertions			
	Japanese		*English*		*Japanese*		*English*		*Japanese*		*English*	
	Error	Percent[1]	Error	Percent	Error	Percent	Error	Percent	Error	Percent	Error	Percent
1	g->n	10.2	ah->ay	4.2	u	6.0	s	5.4	o	5.2	n	18.7
2	g->r	2.1	w->n	4.0	i	3.7	k	4.0	n	4.9	ow	1.7
3	g->y	1.7	ih->t	3.8	k	2.5	ih	2.4	nn	2.6	th	1.4

[1] Percent of total errors

tokens have very limited information to distinguish them from each other. From the analysis in the previous section, this would suggest that recognition is a more difficult task for Japanese with generally shorter words.

By comparing Tables 2 and 3, it is fairly easy to correlate the substitution errors with Japanese word and phone models. For instance, g -> n phone substitution correlates with go -> no word substitution, and g -> y correlates with go -> yon (ranked as the fourth most common word error). However, it is not easy to correlate the English phone and word errors. For instance, ah -> ay phone substitution is caused by substitution of one with nine and five. However, one -> nine is 15th, and one -> five is merely the 29th most common substitution word error. This seems to suggest that the context dependence of phones, which is modeled in word models but not in our phone model, is not as significant for Japanese compared to English. The overall error rate did not differ significantly, however, with an error rate of 2.4% and 5.4% for Japanese and English word models, and 12.9% and 15.1% for Japanese and English phone models respectively

4. Conclusion

We compared the digit recognition task for Japanese and English, using both word models and context - independent phone models. Durations of the recognition units were generally longer and more variable for English. We performed digit recognition tests under the same constraints. Error analysis showed that most errors involved the shorter words, for instance "go" (5) in Japanese, and "oh" (0) in English. Various speech feature vector dimensions, LPC analysis order, and frame durations were tested. Contrary to our expectations, language dependency does not seem to exist in these parameters. For both languages, the feature dimension seems to be optimum at about 16 features. LPC analysis order does not seem to be a crucial parameter. Frame duration seems to be best at 20 msec. Thus, we are optimistic that these parameters need not be re-optimized for each new language, at least for the dedicated task of digit recognition.

We are now in the process of collecting a large Japanese telephone corpus - the Voice Across Japan (VAJ) corpus [6]. This corpus includes both read digit strings and phonetically balanced sentences. Thus, we should be able to compare recognition performance for Japanese and English at a more comparable condition, telephone input via public network. We should also be able to compare results for larger vocabulary, with a more generic task.

5. References

[1]. L. Lamel, J. Gauvain, "Cross-Lingual Experiments with Phone Recognition," Proc. ICASSP 93, April, 1993.

[2]. M. Kojima, J. Picone, and S. Kato, "The Japanese VoiceDialer Database," Texas Instruments Technical Memorandum TRDC-TM-92-01, 1992.

[3]. B. Wheatley, and J. Picone, "Voice Across America: Toward Robust Speaker-Independent Speech Recognition for Telecommunications Applications," Digital Signal Processing, vol. 1, no. 2, pp. 45-63, April, 1991.

[4] G. R. Doddington, "Phonetically Sensitive Discriminants for Improved Speech Recognition," Proc. ICASSP 89, May, 1989.

[5] T. Kobayashi, S. Itahashi, S. Hayamizu, and T. Takezawa, "ASJ Continuous Speech Corpus for Research," Journ. Acoustical Soc. of Japan, vol. 48, no. 12, 1992.

[6]. T. Staples, J. Picone, K. Kondo, and N. Arai, "The Voice Across Japan Database - The Japanese Language Contribution to POLYPHONE," Proc. ICASSP 94, April, 1994.