1—Q—10

A COMPARATIVE ANALYSIS OF JAPANESE AND ENGLISH DIGIT RECOGNITION*

○*Kazuhiro Kondo*[1], *Joseph Picone*[2], *and Barbara Wheatley*[2]

Texas Instruments, [1]Tsukuba Research and Development Center, [2]Central Research Laboratories

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. We will attempt a comparison of Japanese and English recognition with a specific, practical task: digit recognition. In this paper, we will try to analyze the language dependent aspects of this task through experimentation. The recognition performance of both languages will be presented along with a sensitivity analysis of several key recognition parameters.

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, and two American English corpora. 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). 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 [1], which was collected over public telephone lines and thus included considerably more channel noise, as well as background noise. All of the above corpora contain read speech relevant to voice dialing applications, but only the digit strings were used for our comparative analysis.

Analysis of the words in the Japanese corpora showed that 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.

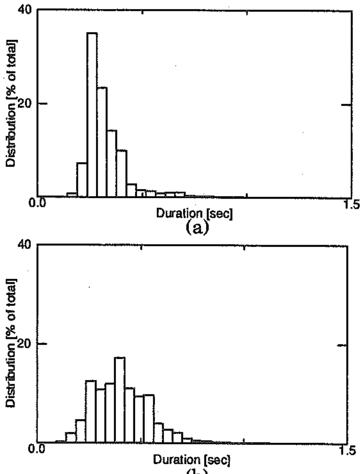

Fig. 1. Distribution of Word Duration in (a) Japanese and (b) English

3. RECOGNITION SYSTEM DESCRIPTION

The recognition system used here is an LPC-based HMM recognizer [2]. Speech is sampled at 8 kHz, LPC analysis is applied, and the LPC parameters are transformed into a feature vector whose components are orthonormal. A unimodal Gaussian continuous distribution model is used along with a Viterbi-style maximum likelihood path scoring in the HMM model.

For both languages, 80% of the speakers were designated as training sets. Of the remaining 20% of the speakers, the pure digit sentences were used for the recognition tests.

4. 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.

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 Japanese 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.

5. 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

*日英両言語における数字認識の比較

近藤 和弘[1]、 ジョー ピコネ[2]、 バーバラ ウィトリー[2]

テキサス インスツルメンツ、

[1] 筑波研究開発センター、[2] 中央研究所

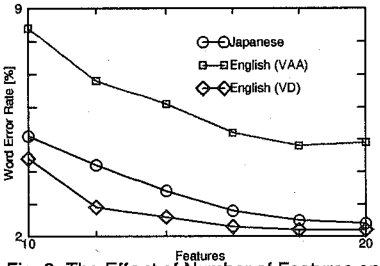

Fig. 2. The Effect of Number of Features on the Recognition Performance

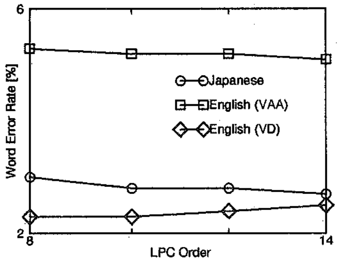

Fig. 3. The Effect of LPC Analysis Order on the Recognition Performance

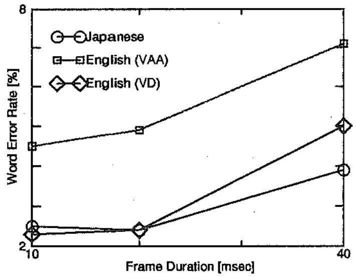

Fig. 4. The Effect of Frame Duration on Recognition Performance

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 VAA corpus and the 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.

5.1. 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 an LPC analysis order of ten. For the Japanese and English VAA corpus, 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 corpus, 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.

5.2. 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).

5.3. Frame Duration

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

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).

6. CONCLUSION

We compared the digit recognition task for Japanese and English. An analysis of the corpora showed that durations of the recognition units were generally longer and more variable for English. We performed digit recognition tests under the same constraints. 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 [3]. With this corpus, 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.

References

[1]. 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.

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

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