RECOGNITION OF CONNECTED DIGIT SPEECH IN JAPANESE COLLECTED OVER THE TELEPHONE NETWORK

Hisashi Kawai and Norio Higuchi

KDD R&D Laboratories Inc.,
2-1-15 Ohara Kamifukuoka-shi, Saitama 356-8502, Japan
{kawai,higuchi}@lab.kdd.co.jp

ABSTRACT

Recognition of connected digits is an important element for applications of speech recognition over the telephone network. This paper describes experimental results on whole word HMM-based speech recognition of connected digits in Japanese. The training data comprises 756860 digits (91.17 hours) uttered by 1963 speakers, while the testing data comprises 304212 digits (31.39 hours) uttered by 852 speakers. The best performance was a word error rate of 0.42% for known length strings obtained using context dependent models. The word error rate was measured as a function of the training data size. The result showed that at least 3302 samples per speaker and 344 speakers are necessary and sufficient for context independent training. Error analysis was conducted on a fraction of the population bearing the major part of recognition errors. The results suggested that such speakers arise not simply from speaker characteristics but from a combination of speaker characteristics and environmental conditions.

1. INTRODUCTION

Recognition of connected digits is an important element for applications of speech recognition over the telephone network, such as voice dialing by number, PIN (personal identification number) entry, and credit card validation.

Numerous studies have been made on the recognition of digits in English [1-3]. In recent years successful results have been obtained by straightforwardly applying HMM-based techniques commonly used in large vocabulary speech recognition. Rabiner et al. showed the effectiveness of whole word HMM for speaker dependent and speaker independent recognition [1]. Wilpon et al. found that incorporating first and second order time derivatives into the acoustic feature vector substantially improves the recognition performance [2]. Buhrke et al. applied VQ-based HMM to telephone speech [3] along with other techniques such as mutual information training, gender modeling, and state number optimization. They achieved a word error rate of minimum 1.7%.

As for Japanese speech, several studies have been published [4-7] after the pioneer studies at the time of DP [8,9]. In spite of those studies we have not seen connected digit recognition adopted in real world applications. We still have to improve the recognition performance and the robustness, and then provide a reliable evaluation of the performance in real operating conditions.

The purpose of this paper is twofold. First, we present our best recognition performance. Secondly, we investigate two factors affecting the recognition performance: training data size, and a small fraction of the population bearing a major part of recognition errors. The latter factor is generally known as the "sheep and goats" problem [10]. In this paper, however, we call them low performance speakers to clarify whether the sheep or the goats are the problem.

This paper is organized as follows. In section 2, we describe a database of connected digit speech collected over the telephone network. In sections 3 and 4, we describe the common experimental set-up, viz. feature extraction and whole word HMM's. In section 5, we show our best recognition performance at present. In section 6, we present the impact of training data size on the recognition performance. In section 7, we analyze the low performance speakers. In section 8, we summarize the paper.

2. TELEPHONE SPEECH DATABASE

The telephone speech database used for HMM training and speech recognition is outlined in Table 1. Each sentence for utterance comprises four connected digits. The speakers of the training data were assigned 100 strings in a written form from a randomized list of numbers 0000-9999, while the speakers of the testing data were assigned 10 strings from the same list. The speakers were requested to pronounce digits as specified in Table 2. Although the speakers were requested not to insert pauses between digits as long as possible, 21% of the utterances were found to contain pauses.

Table 1: Outline of the database for training and testing. The duration does not include sentence initial and final pauses.

	Gender	Number of Speakers	Number of Utterances	Duration (hours)
Training	**Male**	1375	132383	63.50
	Female	588	56832	27.67
Testing	**Male**	428	37038	15.13
	Female	424	39015	16.26

Table 2: Pronunciation of digits. /H/ means vowel elongation.

Digit	Pronunciation	Digit	Pronunciation
0	/zero/	5	/go/, /goH/
1	/ichi/	6	/roku/
2	/ni/, /niH/	7	/nana/
3	/saN/	8	/hachi/
4	/yoN/	9	/kyuH/

The speakers were collected mainly from the Tokyo and Osaka areas. Their ages were evenly spread over the range from 15 to 60 years old. There was no overlap of speakers between the training data and the testing data.

The speakers of the training data uttered in one of the three environments, viz. home, office, or pay phone, while the speakers of the testing data uttered twice in two weeks in all of the five environments, viz. home, office, PHS (personal handy phone system), and cellular phone. The speakers were requested not to use a cordless phone handset at home. The average SNR of recorded speech ranged from 30 to 38 dB depending on the environment.

The data acquisition equipment was connected to the telephone network through an ISDN channel. Analog components affecting the acoustic features of speech are thus limited to the telephone terminal and the subscriber line. The speech was digitized at 8 kHz in 8 bit μ-law format.

3. FEATURE EXTRACTION

The speech was pre-emphasized with a coefficient of 0.95 and windowed with a 32ms Hamming window at every 10ms. The windowed speech was parameterized into a 39 component vector consisting of 12 MFCC's (mel-frequency cepstrum coefficients), their first and second order time derivatives, and first and second order time derivatives of log-energy. Cepstrum mean normalization (henceforth CMN) was applied for channel equalization. A white Gaussian noise less than the quantization level was added before the windowing so that absolutely zero frames do not affect the whole utterance through CMN especially for the lower order MFCC's.

4. WHOLE WORD MODELING

Context independent (henceforth CI) HMM's and context dependent (CD) HMM's were trained using a standard Baum-Welch maximum likelihood estimation. Male speech and female speech were modeled separately. All of the training data shown in Table 1 was used for the modeling. All training and testing were conducted using the HTK HMM toolkit [11].

Table 3 shows the number of states in the case of CI models, which were obtained as a result of optimizing the sentence error rate using the hill-climbing method. The initial values were determined by assigning 3 states to a phoneme. The pause model was not optimized. The states in an HMM were organized in a left-to-right network without skip transitions. The output distribution of each state is a mixture of 15 (for digits) or 60 (for pause) Gaussian distributions. The total number of states and Gaussian distributions were 127 and 1950, respectively.

In the case of CD models, preceding and succeeding environments were grouped into predefined 6 and 5 classes, respectively. The middle states, viz. the states except for the initial two states and the final two states, were shared among the models for the same digit. Consequently 1287 free states consisting of 19350 Gaussian distributions were used in total. The transition matrix was shared among the models for the same digit.

Table 3: Number of HMM'states (lower row) for each digit (upper row).

0	1	2	3	4	5	6	7	8	9	Pause
16	9	14	11	11	13	16	14	12	10	1

5. RECOGNITION RESULTS

The trained models were evaluated on the testing data shown in Table 1. Only the speech from the first session was used. Viterbi decoding was executed without a beam search. The grammar network permitted an arbitrary insertion of pauses into sentence initial, word medial, and sentence final positions. When a speech of unknown gender was recognized, the network of male HMM's and the network of female HMM's were searched in parallel.

The recognition results for CI models and CD models are shown in Tables 4 and 5, respectively. The word error rate for a known length string is defined as $100(D+S)/N$, where D is the number of deletions, S is the number of substitutions, and N is the total number of digits. On the other hand, the word error rate for an unknown length string is defined as $100(D+S+I)/N$, where I designates insertion errors. The difference between the scores for known length strings and for unknown length strings mainly arose from insertion errors.

Table 4: Word error rate for CI models. KL: known length, UL: unknown length. The cellular phone environment is excluded from the mean.

Gender		Environments					
		Home	Office	Pay	PHS	Cel.	Mean
Male	KL	0.47	0.33	0.69	0.58	0.98	0.52
	UL	2.64	3.56	3.23	2.45	4.50	2.99
Female	KL	0.38	0.27	0.85	0.55	1.21	0.51
	UL	2.23	3.39	2.86	2.24	4.43	2.69
Unknown	KL	0.43	0.32	0.78	0.52	1.10	0.51
	UL	2.41	3.31	2.94	2.27	4.36	2.74

Table 5: Word error rate for CD models.

Gender		Environments					
		Home	Office	Pay	PHS	Cel.	Mean
Male	KL	0.36	0.26	0.59	0.47	0.79	0.42
	UL	2.54	3.53	3.26	2.23	4.54	2.91
Female	KL	0.33	0.20	0.84	0.36	1.15	0.43
	UL	2.29	3.75	3.28	2.18	4.52	2.89

6. TRAINING DATA SIZE AND RECOGNITION PERFORMANCE

The training data is the most crucial factor to the recognition performance, while building a speech database is an expensive task. For this reason it is interesting to know how much data is necessary. To investigate the relationship between the training data size and the recognition performance, training with

reduced data was conducted. The number of speakers and the number of utterances per speaker were both reduced to $1/2^n$. Since the data from female speakers is 1/3 that from the male speakers, the experiment was conducted only on the male speakers.

Figure 1 shows the result for CI models. We can see in this figure that the error rate improves slowly when it falls below 0.6%. The boundary of the region where the error rate is smaller than 0.6 indicates that the number of speakers tends to have a greater effect on the error rate than the number of utterances per speaker, which coincides with Shirotsuka's report [4] although the error rates are substantially better in our case. If we regard the error rate of 0.6% as the minimal error rate achieved by all the available training data, the necessary conditions for the best error rate are as follows:

1. The number of speakers is greater than 344.
2. The number of utterances is greater than 8256. Namely, the number of samples per digit is greater than 3302.

Figure 2 shows the result for CD models. In this case the improvement of the error rate becomes smaller when it is below 0.45%. If we regard the error rate of 0.45% as the minimal score, the necessary condition for the best score is as follows:

- The number of utterances is greater than 16512. Namely, the number of samples per digit is greater than 6605.

No apparent requirement in terms of the number of speakers is observed in Figure 2. Since 30 HMM's correspond to a digit in the case of CD modeling, the number of samples necessary for an HMM is 220.

7. ANALYSIS OF LOW PERFORMANCE SPEAKERS

It is well known that recognition errors are not distributed equally over the population, but tend to be concentrated in a fraction of the population [10].

The first question is how many low performance speakers exist. Figure 3 shows average word errors rate for 10% fractions of the speakers. The speakers are sorted by word error rate. The HMM's used were CI models trained with the male speech. The testing data uttered in the first session was used for the recognition. Figure 3 indicates that the recognition errors are

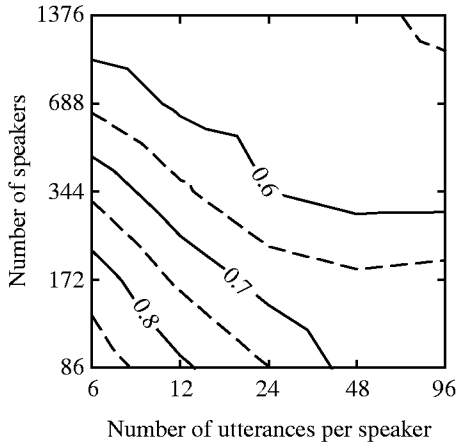

Figure 1: Relationship between training data size and word error rate (%) for CI models.

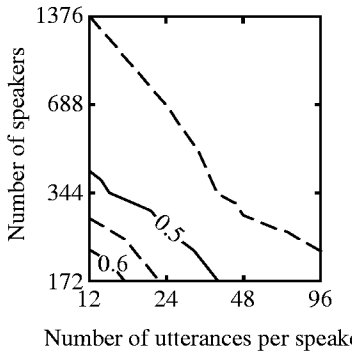

Figure 2: Relationship between training data size and word error rate (%) for CD models.

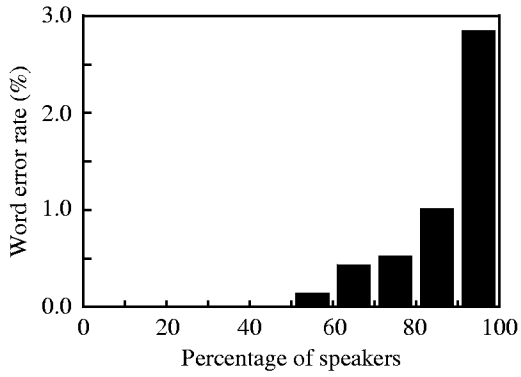

Figure 3: Word error rates for 10% fractions of the speakers.

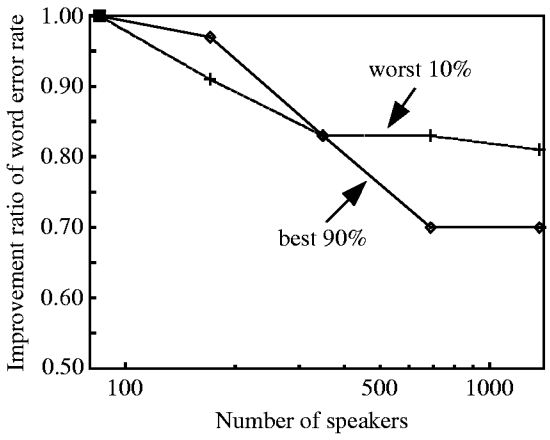

Figure 4: Improvement ratio of word error rate as a function of the number of speakers (96 samples/speaker).

concentrated in the worst half, and especially in the worst 10% of the population. The worst 10% of the population accounts for 57% of all the recognition errors.

The second question is what we can do to decrease the number of low performance speakers. Is it possible to decrease it if we increase the training data size? Figure 4 shows the improvement ratio of word error rate for CI models as a function of the number of speakers. The improvement ratio is defined as an error rate normalized by the value for 85 speakers. This figure indicates that increasing the training data size is effective as long as the number of speakers is less than 688 (for the best 90% speakers) or 344 (for the worst 10% speakers). Although the improvement ratio for the worst 10% speakers further decreases, the improvement is trivial. It is also noteworthy that the improvement for the worst 10% speakers is slower than that for the best 90% speakers. The result obtained here suggests that augmenting the training data is an inefficient means for decreasing the number of low performance speakers.

The third question is whether or not the recognition errors of the low performance speakers are caused only by speaker characteristics. To answer this question, the recognition errors of the 42 low performance speakers were analyzed individually. It was found that 41 out of the 42 speakers did not have a recognition error in at least one of the four environments. This result suggests that the reason for the low performance speakers is a complex of speaker characteristics and environmental conditions including stochastic factors such as ambient noise.

8. SUMMARY

In this paper we have presented experimental results on recognizing connected digit speech in Japanese spoken over the telephone network. The standard methods were used for feature parameterization and modeling digits. The acoustic feature vector comprised of 12^{th} order MFCC's, their first and second time derivatives, and the first and second time derivatives of log-energy. Digits were modeled by whole word HMM's with continuous output distribution. The training data comprised 756860 digits (91.17 hours) uttered by 1963 speakers, while the testing data comprised 304212 digits (31.39 hours) uttered by 852 speakers.

The word error was measured as a function of the training data size. The results showed that at least 3302 samples per speaker and 344 speakers are necessary for the CI modeling, while 6605 samples were necessary for the CD modeling. In other words, a CI HMM requires 3302 samples, while a CD HMM requires 220 samples. This discrepancy should be studied in future work.

Analysis was conducted on a fraction of the population bearing the major part of recognition errors. The worst 10% fraction was found to account for 57% of recognition errors. The experiment on training data size and recognition performance suggested that the number of these low performance speakers cannot be decreased by augmenting the training data. It was also found that speaker characteristics create low performance speakers in combination with environmental conditions. The mechanism of how they combine with each other will be studied in future work.

REFERENCES

1. Rabiner, L.R., Wilpon, J.G., and Soong, F.K. "High performance connected digit recognition using hidden Markov models," *IEEE Trans. ASSP, Vol. 37, No. 8, 1214-1225,* 1989.

2. Wilpon, J.G., Lee, C.-H., and Rabiner, L.R. "Connected digit recognition based on improved acoustic resolution," *Computer Speech and Language, Vol. 7, No. 1, 15-26,* 1993.

3. Buhrke, E.R., Cardin, R., Normandin, Y., Rahim, M., and Wilpon J. "Application of vector quantized hidden Markov modeling to telephone network based connected digit recognition," *Proc. ICASSP 94, I-105,* 1994.

4. Shirotsuka, O., Arima, I., and Kawai, G. "Spoken Japanese Sentence Recognition Based on Hidden Markov Models," *Reports of Autumn Meet., Acoust. Soc. Jpn., 1-5-16, pp. 31-32,* 1991.

5. Isobe, T. and Murakami K. "Nationwide collection of telephone speech and evaluation of recognition using the data," *Reports of Spring Meet. Acoust. Soc. Jpn., 2-Q-26, pp. 135-136,* 1993 (in Japanese).

6. Kondo, K., Picone, J., and Wheatley, B. "A Comparative analysis of Japanese and English digit recognition," *Proc. ICASSP 94, I-101,* 1994.

7. Matsuoka, T., Uemoto, N., Matsui, T., and Furui, S. "Acoustic modeling for connected digit speech recognition," *IEICE Technical Report, SP95-23, pp. 39-44,* 1995 (in Japanese).

8. Sakoe, H. "Two-Level DP-Matching – A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition," *IEEE Trans. ASSP, Vol. ASSP-27, No. 6, pp.588-595,* 1979.

9. Hataoka, N., Asakawa, Y., and Komatsu, A. "Speaker-Independent Connected Digit Recognition," *Proc. ICASSP, 35.1.1-4,* 1984.

10. Doddington, G.R., and Schalk, T.B. "Speech recognition: turning theory to practice," *IEEE spectrum, Vol. 18, No. 9, pp. 26-32,* 1981.

11. Young, S.J., Woodland, P.C., and Byrne, W.J. "HTK-Hidden Markov Model Toolkit, Version 1.5," Cambridge University Engineering Department and Entropic Research Laboratories Inc., 1993.