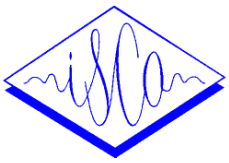


2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Performance of Speaker-Independent Japanese Recognizer as a Function of Training Set Size and Diversity

O. Shirotsuka, G.Kawai**, M. Cohen**and J. Bernstein***

*NTT Data Communications Systems Corporation
Kowa-Kawasaki Nishi-Guchi Bldg.,
66-2 Horikawa-cho, Saiwai-Ku
Kanagawa, 210 Japan

**SRI International
333 Ravenswood Avenue, Menlo Park,
California, 94025 U.S.A

Abstract

Experiments investigated the effects of training set size and diversity of speech data in training an HMM-based, speaker-independent, continuous Japanese speech recognition system. Two different types of diversity were investigated: speaker diversity and phonetic diversity. The results indicate that greater amounts of training data improve recognition performance and that, given a fixed amount of training data, greater diversity of training materials both in terms of speakers and phonetic contexts improve recognition performance.

10.21437/ICSLP.1992-65

1 Introduction

Many new speech data collection projects are starting in the United States and in other parts of the world. One aspect of the database design revolves around questions of diversity: How many different utterances? Of what linguistic material? From how many different people?

This paper provides some recognition performance data that helps answer these questions. In particular, the paper presents several experiments that tried to identify efficient methods for using speech data to train an HMM-based speech recognition system. The system used in these experiments was developed at SRI International [1] and then adapted for use in Japanese as part of a cooperative research effort between NTT-Data and SRI International. System development was focused on recognition of utterances used for scheduling conference rooms. This paper first describes the design and collection of spoken training material for these tasks, and then describes an experiment that examined the performance tradeoffs from training on greater and lesser amounts of speech from smaller and larger numbers of speakers.

A second set of experiments was run to examine the relationship between phonetic diversity in the training set and recognition performance. These experiments involved measuring recognition performance as we varied the number of different sentences (i.e., word strings) and the number of different triphones.

2. Task Domain and Materials

We chose a conference room scheduling (CRS) task for design and evaluation of the system. The task domain was based on an imaginary computer-based system that is used to schedule conference rooms in office.

A person familiar with reserving conference rooms invented 840 different sentences in this domain. Examples of the types of sentences included in this task domain might be translated as:

[example 1]
木曜日の午前中に使える会議室はありますか
<mukuyoo no gozen chuu ni tukaeru kaigisitu wa arimasuka>
(Is there a conference room available on Tuesday morning?)

[example 2]
佐藤さんがいらっしゃる会議室はどれですか
< sato san ga irassyaru kaigisitu wa dore desuka>
(Which conference room is Mr. Sato in now?)

The sentences were chosen primarily to include as many different verb forms and syntactic forms as possible, with an eye toward exercising and refining a natural language parser for Japanese.

We collected speech data from 110 people, using a high-quality noise-cancelling close-talking microphone in a quiet room. Each person read 84 sentences, and

thus we obtained 11 tokens for each sentence type and 9240 utterances in total. The vocabulary size of the CRS corpus is 158, and average length of CRS sentences is 11.6 words.

3. Recognition System

The research presented here used a speaker independent, continuous speech recognition system with context-dependent phone models, including word-specific, triphone, generalized-triphone, biphone, and generalized biphone. These models are smoothed together, along with context-independent models, using the deleted interpolation algorithm [2]. This smoothing maintains robustness even for highly specific contexts with small numbers of training examples.

In this system for recognition of Japanese, pronunciations were represented with an inventory of 43 phoneme-level models and words were modeled as linear sequences of phoneme-level models. The system was run in all the following experiments with no grammar, thus the perplexity of the task is equal to the vocabulary size (158).

Some detail aspects of the recognition system design are presented in Table 1.

States Per Phone Model	3
Sampling Frequency	16Khz, 16bit
Analysis Condition	Hamming, Frame Length 25.6ms
Acoustic Features	12-Dimentional Mel-Cepstrum 12-Dimentional Δ Mel-Cepstrum Log Energy Δ Log Energy
Vector Quantization	LBG Algorithm, Size = 256

Characteristics of the recognition system

Table 1

4. Experiment 1: Accuracy as a function of training size and speaker diversity

The first experiment run in the conference room scheduling (CRS) domain tested recognition performance as a function of which subset of the total speech data was used in training. Although there have been many reports on the effect of training material on recognition performance, there have not been many studies on how the number of speakers affects recognition performance, given the same amount of data [3].

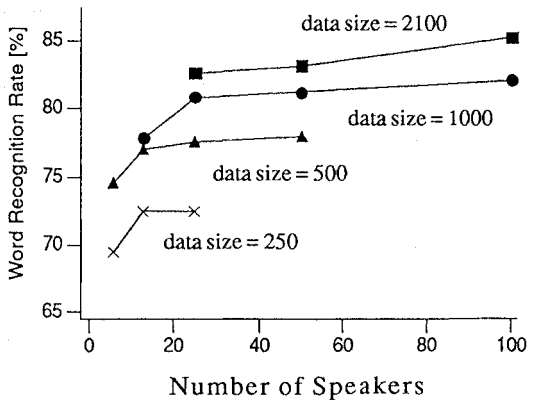
In our first experiment, we observed recognition performance while varying both the number of speakers and the number of utterances used from each speaker. We measured recognition performance while holding the number of training utterances constant at several levels (60, 130, 250, 500, 1000, 2100, 4200, 8400 sentences), and varying the number of different speakers.

For test data, we randomly selected 4 male speakers speaking 84 sentences each. Table 2 and Figure 1 show experimental conditions and results. Figure 1 shows how total data size and the number of speakers for a given data size affect recognition performance. Figure 1 shows not only the general tendency toward performance improvement with an increase in the total amount of data, but also the general tendency for improvement from an increase in the number of speakers with a fixed size of training data. This tendency can be observed in the upward trend of lines that express the performance at each fixed data size.

	X = Number of Speakers				
Y=Number of Sentences	X = 6	X = 13	X = 25	X = 50	X = 100
Y = 10	52.0	65.0	72.4	78.0	82.1
Y = 21	61.2	72.4	77.6	81.2	84.0
Y = 42	69.4	77.0	80.9	83.2	85.3
Y = 84	74.6	77.9	82.7	84.8	85.6

Recognition accuracy (in percent word correct) for a range of numbers of speakers and numbers of sentences per speaker.

Table 2



Recognition accuracy (in percent word correct) for different numbers of speakers for a range of data set sizes.

Figure 1

The two principal results are: (1) a larger number of training utterances generally increases performance, and (2) for the same number of training utterances, more speakers speaking fewer utterances is better than fewer speakers speaking more utterances. For example, as shown in Table 2, 100 speakers speaking 10 utterances each gives better results than 13 speakers speaking 84 utterances each, even though the total number of utterances is almost 10 percent larger in the latter case. We had hypothesized that for the most difficult test speaker, the recognition accuracy might be even more sensitive to training set diversity than for the average speaker in the test set, but the improvement given greater speaker diversity in the training set was about the same for the worst speaker as the average.

5. Experiment 2 and 3: Accuracy as a function of phonetic diversity

The goal of the experiments described in this section was to examine recognition performance as a function of phonetic diversity of the training set. Related work by Hon [3] showed improved recognition performance on vocabulary-independent recognition in English with increased triphone coverage.

The same (CRS) database was used as described above. In experiment 2, recognition performance was measured in four different experimental conditions, based on four different training sets. Each training set had the same number of speakers (80) and the same number of utterances by each speaker (8), resulting in a total of 640 utterances. The diversity of phonetic material in each training set was varied as follows:

training set 1
Each of the 80 speakers spoke a different set of 8 sentences, resulting in 640 different sentences in the training set.

training set 2
Each of 320 sentences were spoken by 2 different speakers.

training set 3
Each of 128 sentences were spoken by 5 different speakers.

training set 4
Each of 64 sentences were spoken by 10 different speakers.

The same test set, consisting of 1680 utterances by 20 speakers, was used for each test. Each of the 20 test speakers said 84 sentences. There were 168 different sentences in the test set, each one spoken by 10 speakers. No sentences in the test set occur in any of the training sets. No speakers in the test set occur in any of the training sets. The test set contained 16102 words, with 144 unique words.

Table 3 shows percent word correct for the four tests, as well as counts of the number of different triphones, different sentences, and different words occurring in the training set, and the number of unique words in the test set that do not occur in the training set.

The results show improved performance as the phonetic diversity of the training material increases. The differences in recognition results between all pairs of tests are significant at the 99.9% level except for training set 2 vs training set 3, which was not significant at the 95% level. As shown in Table 3, for these cases, the number of unique triphones and words, as well as the coverage of words in the test set, gets higher as the number of unique sentences gets higher.

	training set			
	1	2	3	4
recognition perf.	76.8	75.3	75.2	72.3
unique triphones	1093	1000	832	678
unique sentences	640	320	128	64
unique words	155	154	145	120
missing test words	0	0	7	26

Training set characteristics and recognition performance (in percent word correct) for different sentence diversities.

Table 3

	training set	
	1	2
recognition perf.	80.4	77.7
unique triphones	1055	889
total triphones	22716	22665
unique sentences	294	294
unique words	155	148
missing test words	0	6

Training set characteristics and recognition performance (in percent word correct) for different diversities.

Table 4

In experiment 3, recognition performance was measured in two experimental conditions, based on two training sets. The goal was to hold the number of unique sentences constant, while varying the number of unique triphones covered by the training set. Each training set had the same number of speakers (70), same number of utterances by each speaker (42), and the same number of unique sentences (294). Each training set contained a total of 2940 utterances.

The sentences in training set 1 were chosen to maximize the number of different triphones and the sentences in training set 2 were chosen to minimize the number of unique triphones, while attempting to keep the amount of training material roughly the same in the two cases. The test set, for both cases, consisted of 2520 utterances (252 different sentences) by 30 speakers. No sentences or speakers in the test set were in either of the training sets.

Table 4 shows percent word correct for the two tests, counts of the number of different triphones, number of total triphones, number of different sentences and different word occuring in the training set, and the number of unique words in the test set that do not occur in the training set. for each training set.

The results show improved performance with increased triphone diversity in the training set. The difference is significant at the 99.9% level. The total number of triphones in each case is almost the same, suggesting roughly the same amount of training material, while the number of unique triphones varies by almost 20%.

6. Conclusions

Our results show improved recognition performance with larger training sets, and, for a given training set size, with increased diversity of training set material, both in terms of speaker diversity and phonetic diversity. One novel contribution of the work presented here is the density of performance data points as a function of training set size and number of speakers. Several researchers have extrapolated from a single suggestive result reported for the DARPA Resource Management data base that large amounts of speech data from a few speakers is as good as a similar total amount of data collected from many speakers. Our results suggest that this extrapolation may be questionable and that for the time being speech data collection efforts need to face the cost and inconvenience of recording many speakers. In addition, our results suggest that improved recognition can be achieved by varying the sentence material in a database, rather than using the same material with different speakers.

7. Acknowledgement

The authors wish to thank the members of the speech group of NTT Data Communications Systems Corporation and the members of the Speech Research and Technology Program of SRI International.

References

1) Cohen, Murveit, Bernstein, Price & Weintraub, "The DECIPHER Speech Recognition System", in proceedings of ICASSP-90, 1990, pp. 77-81.

2) F. Jelinek and R. L. Mercer, "Interpolated estimation of markov source parameters from sparse data", in Pattern Recognition in Practice, E. S. Gelsema and L. N K anal, Eds. Amsterdam: North-Holland, 1980, pp. 381-397.

3) Kubala and Schwartz, " A New Paradigm For Speaker Independent Training",in proceedings of ICASSP-91, 1991, pp. 833-836.

4) Hon, "Vocabulary-Independent Speech Recognition: The VOCIND System", PhD thesis, Carnegie-Mellon University, March 1992.