THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.28, No.2E 2009. 6. pp. 66-71

The Effect of the Number of Training Data on Speech Recognition

Chang-Young Lee*
*Div. of Information System Engineering, Dongseo University
(Received April 30 2009; accepted May 22 2009)

Abstract

In practical applications of speech recognition, one of the fundamental questions might be on the number of training data that should be provided for a specific task. Though plenty of training data would undoubtedly enhance the system performance, we are then faced with the problem of heavy cost. Therefore, it is of crucial importance to determine the least number of training data that will afford a certain level of accuracy. For this purpose, we investigate the effect of the number of training data on the speaker-independent speech recognition of isolated words by using FVQ/HMM. The result showed that the error rate is roughly inversely proportional to the number of training data and grows linearly with the vocabulary size.

Keywords: *Speech recognition, Number of training data, FVQ, HMM*

I. Introduction

As a method of communication between man and machine, speech recognition provides a very effective interface. Speech input to a machine is about twice as fast as information entry by a skilled typist [1]. The earliest attempt to devise systems for automatic speech recognition by machine is traced back to 1952 when the researchers at Bell Laboratories built a system for isolated digit recognition for a single speaker [2]. Since then, lots of endeavors have been made for over five decades to enhance the recognition accuracy.

Depending on the system performance, the perception of the speakers is known to be divided in two categories [3]. When the recognition error rate is, say 5%, then the system makes an error, on average, once in 20 tries and in this case the user tends to attribute the error to an improper and/or uncooperative speaking mode on his (or her) own part. It is interesting to note that the absolute level of performance is relatively unimportant so long as the recognition accuracy exceeds a certain level. This means that the difference between accuracies of 95% and 98%, for example, is insignificant to the user. If the recognition accuracy falls below some level, say 90%, on the other hand, the perception of the user is that the system makes too many errors and is therefore unreliable.

Since so many factors are involved in the procedures for speech recognition, it is not easy to isolate the effects of some parameters from the others. However, it is beyond all question that more training data would yield better performance in speaker-independent speech recognition. Plenty of training data will afford the system the chance of being exposed to more diverse patterns. More trained, more robust. If the training samples could be provided to our content and the system is allowed to be trained accordingly, then the situation approaches the speaker-dependent case which yields in general much

Corresponding author: Chang-Young Lee (seewhy@dongseo.ac.kr)
Div. Of Information System Engineering Jurye San 69-1, Sasang, Pusan 617-716, Korea

superior accuracy to the speaker-independent one.

Though it is desirable to have as many training data as possible, such a scheme raises the problem of various costs including time and memory. Sometimes, having enough training data is even impractical. There have been lots of efforts [4-13] to overcome the problem of insufficient amount of training data, but such solutions cannot be cure-all.

For these reasons, there should be a compromise between the two goals of better performance and less cost. Instead of estimating the required number of training data by rule of thumb, it is advisable to have a criterion that helps in determining the economical number of data that would provide a certain level of recognition accuracy. Keeping this in mind, we investigate the following in this paper:

(1) The number of training data vs. recognition error rate in speaker-independent speech recognition for various vocabulary sizes
(2) Mathematical modelling and numerical analysis for the results of (1).
(3) Estimation of the required number of training data that would give specific levels of recognition accuracy

II. Experiment

Our experiments were performed on a set of phone-balanced 350 Korean words. In order to study the effect of the vocabulary size, the words were divided into four sets as follows. The sets A, B, and C are disjoint each other and D is the union of those three.

Forty people including 20 males and 20 females participated in speech production. Speech utterances of them were divided into three disjoint groups as follows.

Among the 34 people's speeches of the group I, P people's speeches were used in codebook generation and training of the system. The value of P was changed from 4 to 34 in steps of 2 and thus the recognition performance as a function of the number of training data has been investigated.

The system parameters were updated on each iteration of training. In order to choose which values of parameters to use in actual test of speech recognition, some test speeches are necessary. The parameters that yield the best performance on the group II were stored and used for the group III to obtain the final performance of the speaker-independent speech recognition system. This prescription prevents the system from falling too deep into the local minimum driven by the training samples of the group I and hence becoming less robust against the speaker-independence when applied to the group III.

The speech utterances were sampled at 16 kHz and quantized by 16 bits. 512 data points corresponding to 32 ms of time duration were taken to be a speech frame for short-term analysis. The next frame was obtained by shifting 170 data points, thereby overlapping the adjacent frames by $\fallingdotseq 2/3$ in order not to lose any information contents of coarticulation. To each frame, the Hanning window was applied after pre-emphasis for spectral flattening. MFCC feature vectors of order 13 were then obtained for subsequent processing.

Codebooks of 512 clusters were generated by the Linde-Buzo-Gray clustering algorithm on the MFCC feature vectors obtained from the speeches of P persons from the group I of Table 2. The distances between the vectors and the codebook centroids were calculated and sorted. Appropriately normalized fuzzy membership values were assigned to the nearest two clusters and a train of two doublets (cluster index, fuzzy membership) fed into HMM for speech recognition.

For the HMM, a non-ergodic left-right (or Bakis)

Table 1. Four sets of speech data divided for studying the effect of the vocabulary size.

Set ID	Number of Words
A	50
B	100
C	200
D	350

Table 2. Division of the 40 people of speech production into three groups.

Group ID	Number of People
I	34
II	2
III	4

model was adopted. The number of states that is set separately for each class (word) was made proportional to the average number of frames of the training samples in that class [14]. Initial estimation of HMM parameters $\lambda=(\pi,A,B)$ was obtained by K-means segmental clustering after the first training. By this procedure, convergence of the parameters became so fast that enough convergence was reached mostly after several epochs of training iterations.

Backward state transitions were prohibited by suppressing the state transition probabilities a_{ij} with $i>j$ to a very small value but skipping of states was allowed. The last frame was restricted to end up with the final state associated with the word being scored within a tolerance of 3.

Parameter reestimation was performed by Baum-Welch reestimation formula with scaled multiple observation sequences to avoid machine-errors caused by repetitive multiplication of small numbers. After each iteration, the event observation probabilities $b_i(j)$ were boosted above a small value [15].

Three features were monitored while training the HMM parameters: (1) the recognition error rate for the group II of Table 2, (2) the total probability likelihood of events summed over all the words of the training set according to the trained model, and (3) the event observation probabilities for the first state of the first word in the vocabulary list. Training was terminated when the convergences for these three features were thought to be enough. The parameter values of $\lambda=(\pi,A,B)$ that give the best result for the group II were stored and used in speech recognition test on the group III.

III. Results and Discussion

Figure 1 shows the recognition error rate E vs. the number of training people P for various vocabulary sizes. It is seen that E decreases rapidly at first and then converges toward a certain level as P is increased.

The effect of increasing the number of training data can be seen clearly from the observation probability density $b_i(j)$. Figure 2 shows the observation probability density $b_1(j)$ for the first state of the first word in the vocabulary list of the set B in Table 1. The abscissa denotes the number of training iterations of HMM. In Figure 2(a) where $P=4$, only a few of $b_1(j)$, $j=1\sim512$, have had chances of learning. In Figure 2(b) where $P=34$, meanwhile, most of $b_1(j)$ have learned something. It is revealed that the contents in the parameters become richer as P increases, the obvious reason being that the system has more chances of learning enriched patterns.

In order to analyze the data in Figure 1 numerically, we first consider the following model:

$$E = E_\infty + a\exp(-bP) \tag{1}$$

$E_\infty \equiv E(\infty)$ represents the error rate when the number of training data is infinite. This model is based on the observation that there's no a priori reason that E falls to zero for large P. Instead, we

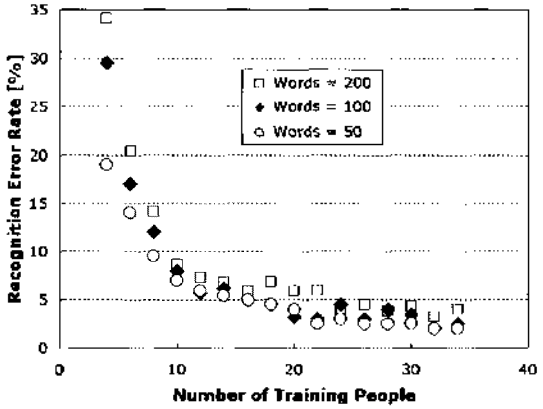


Figure 1. The recognition error rate vs. the number of training people for various vocabulary sizes.

are assuming by Eq. (1) that E converges to a (possibly nonzero) value E_∞ as P goes to infinity. The two parameters a and b are adjustable for the best fit of the model to the data.

The objective function to minimize is given by

$$O = \sum_{i=1}^{N} [E_\infty + a\exp(-bP_i) - E_i]^2 \qquad (2)$$

which is the sum of squares of the deviations between the experimental data and the value calculated from Eq. (1). N is the number of data points (P_i, E_i). $i = 1 \sim N$.

Unless we know E_∞ in Eq. (1) beforehand, it is not easy to decide a and b from curve-fitting. Hence, we should vary E_∞, get a and b therefrom, and calculate the objective function (2). After all, the values of E_∞, a and b are chosen in such a way to minimize O. Figure 3 shows the result for the set C of Table 1, i.e., for the vocabulary size of 200 words.

Since the above result is unsatisfactory especially in the region of small P, we consider another mathematical model:

$$E = \alpha P^\beta \qquad (3)$$

By this model of power law, we are assuming naturally that the recognition error rate converges to zero for sufficiently large training data. α and β are adjustable parameters which should depend on the vocabulary size W. The fitting of Eq. (3) to the data is not difficult and the result is given by Figure 4. Figure 4(b) is the replot of Figure 4(a) by taking the logarithms of both E and P. It might be said that the power law explains the data reasonably well. We note that the slope of the fitted line in Figure 4(b) is close to -1.

The exponent β in Eq. (3) is worth of note. Table 3 shows the values of β for various vocabulary sizes.

It is interesting to see that the exponent of Eq. (3) is close to -1, i.e.,

$$\beta \approx -1 \qquad (4)$$

which means that the recognition error rate is roughly inversely proportional to the number of

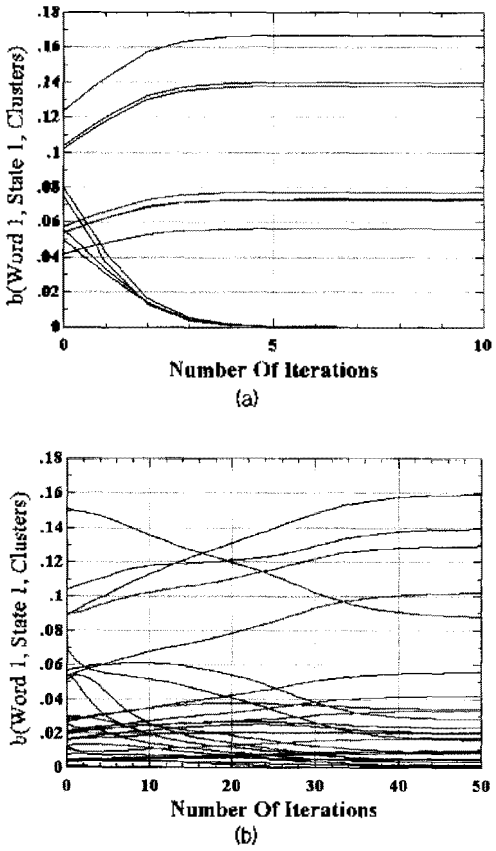


Figure 2. The observation probability density $b_1(j)$ for the first state of the first word in the vocabulary list of the set B in Table 1. The abscissa denotes the number of training iterations of HMM. The number of training people is (a) $P=4$ and (b) $P=34$.

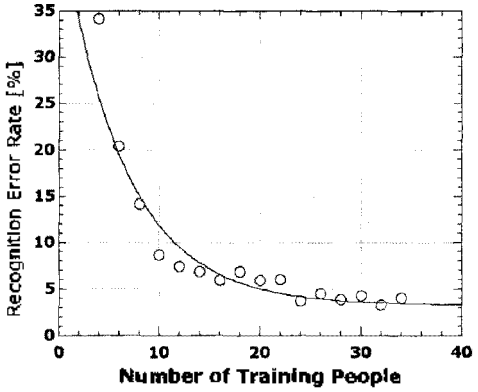
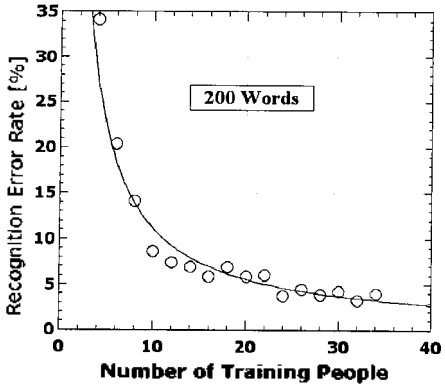
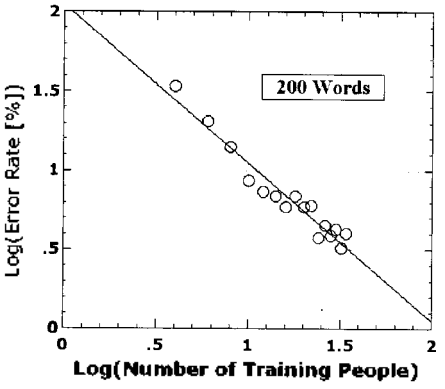


Figure 3. Graph of the recognition error rate vs. the number of training people. The solid line is the result of curve-fitting according to Eq. (1).



(a)



(b)

Figure 4. Graph of the recognition error rate vs. the number of training people for words=200. The solid lines represent the curve-fitting results according to the power law as given by Eq. (3). (b) is the replot of (a) by taking the logarithms of both E and P. The slope of the line in (b) is close to -1.

Table 3. The values of the exponent β in Eq. (3) for various vocabulary sizes.

Number of Words	50	100	200	350
β	-1.08	-1.11	-1.00	-0.98

training data. This result suggests therefore that the number of training people should be doubled if we are to halfen the error rate.

As is explained earlier in this paper, the recognition system is required to surpass some level of recognition accuracy in order to give users the perception of reliability. We examine the necessary number of people to achieve some specific recognition error rates. Figure 5 shows the results obtained from Eq. (3) with the values of α and β obtained by curve-fitting.

We see that the required number of people increases roughly linearly with the number of vocabulary words at least in the regime under our study. Combining this observation with Eqs. (3) and (4), we obtain an empirical law

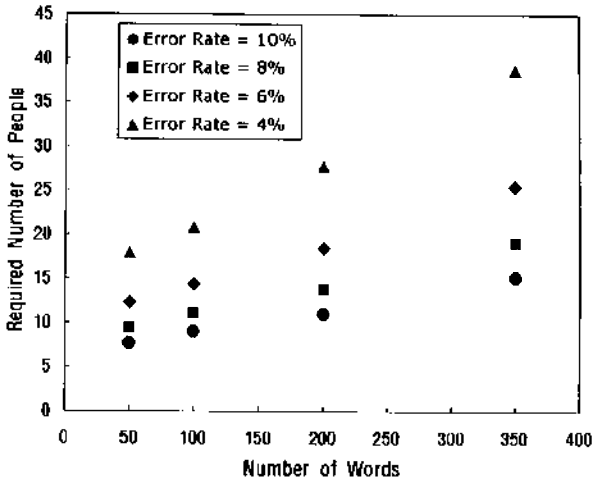


Figure 5. Graph of the required number of people in order to achieve some specified levels of recognition accuracy vs. the vocabulary size.

$$E \propto \frac{W}{P} \tag{5}$$

in the limit of large vocabulary. It might be inferred that, in order to keep a specified level of accuracy, the number of training people should be increased in proportion to the vocabulary size. This behavior implies the difficulty associated with speaker-independent speech recognition of large vocabulary isolated words.

V. Conclusion

In this paper, we studied the performance of the speaker-independent speech recognition of isolated words as a function of the training data and the vocabulary size by using FVQ/HMM. From the curve-fitting of the experimental data, we found that the recognition error rate is roughly inversely proportional to the number of training people. It was also revealed that the required number of training people to achieve a certain level of system performance grows linearly with the vocabulary size. From these

observations, we obtained an empirical law that relates the error rate, the number of training data, and the vocabulary size.

References

1. G. Kaplan, "Words Into Action I," *IEEE Spectrum*, vol. 17, pp. 22–26, 1980.
2. K. H. Davis, R. Biddulph, and S. Balashek, "Automatic Recognition of Spoken Digits," *J. Acoust. Soc. Am.*, vol. 24, no. 6, pp. 637–642, 1952.
3. L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, pp. 485–486, 1993.
4. L. R. Bahl, F. Jelinek, and R. L. Mercer, "A Maximum Likelihood Approach to Continuous Speech recognition," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. PAMI-5, pp. 179–190, 1983.
5. F. Liu, Y. Lee, and L. Lee, "A Direct-Concatenation Approach to Train Hidden Markov Models to Recognize the Highly Confusing Mandarin Syllables with Very Limited Training Data," *IEEE Trans. on Speech and Audio Processing*, vol. 1 no. 1, pp. 113–119, 1993.
6. M. Demirekler, F. Karahan, and T. Ciloglu, "Fusing Length and Voicing Information, and HMM Decision Using a Bayesian Causal Tree Against Insufficient Training Data," *Proc. 15th International Conference on Pattern Recognition*, vol. 3, pp. 102–105, 2000.
7. M. Inoue and N. Ueda, "Exploitation of Unlabeled Sequences in Hidden Markov Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 1570–1581, 2003.
8. V. Siivola and A. Honkela, "A State-Space Method for Language Modelling," 2003 *IEEE Workshop on Automatic Speech recognition and Understanding*, pp. 548–553, 2003.
9. S. Sivadas and H. Hermansk, "On Use of Task Independent Training Data In Tandem Feature Extraction," *ICASSP 2004*, vol. 1, pp. 541–544, 2004.
10. F. Wessel and H. Ney, "Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 1, pp. 23–31, 2005.
11. P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice Modeling with Sparse Training Data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.
12. N. Jakovljevic and D. Pekar, "Description of Training Procedure for AlfaNum Continuous Speech Recognition," *EUROCON 2005*, pp. 1646–1649, 2005.
13. M. Schaffner, S. E. Krüger, E. Andelic, M. Katz, and A. Wendemuth, "Limited Training Data Robust Speech Recognition Using Kernel-Based Acoustic Models," *ICASSP 2006*, vol. 1, pp. 1137–1140, 2006.
14. M. Dehghan, K. Faez, M. Ahmadi, and M. Shridhar, "Unconstrained Farsi Handwritten Word Recognition Using Fuzzy Vector Quantization and Hidden Markov models," *Pattern Recognition Letters*, vol. 22, pp. 209–214, 2001.
15. S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *Bell Systems Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.

[Profile]

• Chang-Young Lee

Chang-Young Lee received the B.S., M.S., and Ph.D. degrees from Seoul National University, KAIST, and State University of New York at Buffalo in 1982, 1984, 1992, respectively.
From March 1984 to August 1988, he was a senior engineer in LG Electronics.
Since March 1993, he has been a professor at the department of Information System Engineering, Dongseo University.
His major research area is speech recognition.