論文 / 著書情報
Article / Book Information

Title	A Training Procedure for Isolated Word Recognition Systems
Author	SADAOKI FURUI
Journal/Book name	IEEE Trans. on ASSP, Vol. 28, No. 2, pp. 129-136
発行日 / Issue date	1980,
権利情報 / Copyright	(c)1980 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

A Training Procedure for Isolated Word Recognition Systems

SADAOKI FURUI, MEMBER, IEEE

Abstract—A procedure has been devised to reduce the amount of training required for a phoneme-based speaker-dependent word recognition system and still maintain performance. Each new speaker is required to provide utterances of only a fraction of the entire vocabulary as a training set. A set of transformation rules is used to estimate phoneme templates for the entire vocabulary from phoneme templates included in the training. The transformation rules are obtained in a pretraining procedure in which a group of speakers provides utterances of the entire vocabulary and multiple regression analysis (MRA) is used to obtain linear estimates of the entire phoneme template set in terms of the set designated as training templates. This group of speakers is generally distinct from the group of training speakers. Thus, since the transformation rules are established independent of the training speakers, the entire procedure can be considered a hybrid speaker-dependent/speaker-independent system. Results of recognition experiments using spoken digits uttered by 30 male and female speakers and 67 airport names uttered by 30 male speakers have ascertained the effectiveness of this training procedure. A mean recognition accuracy of 98.2 percent was obtained for the latter utterance set after a 12-word training procedure.

I. Introduction

RESEARCH on automatic recognition of spoken words has been carried out since about 1950 and, recently, connected speech recognition studies have been made by many investigators. However, research on isolated word recognition is still quite important from the viewpoint of practical use. Presently, some practical isolated word recognition systems have been produced and tested. Most need to be adjusted to each speaker's voice before recognition in order to achieve good performance. For recognition systems with large vocabularies and many speakers, this total amount of training is quite laborious. It would be highly desirable in such situations if training could be carried out using only a fraction of the entire vocabulary or even a short sentence.

Kohda *et al.* [1] investigated a 20-word recognition system using a restricted number of training samples. It was established that if the training words are selected optimally from the 20 words based on the spectral similarity between phonemes of training words and those of nontraining words, the number of training words can be reduced to ten, keeping the recognition rate comparable to that which was obtained in the case when all 20 words were used for training.

The author previously investigated an interspeaker normalization method which adjusts the effects of the glottal wave spectrum and vocal tract length on the speech spectrum. The effectiveness of this method was ascertained by a ten-digit recognition system, but the efficiency was not enough to completely normalize the interspeaker variability [5].

In this paper an improved efficient training method, which uses only a few vocabulary words for training per speaker, is presented.

Manuscript received October 3, 1978; revised July 17, 1979 and October 2, 1979.

The author is with Musashino Electrical Communication Laboratory, Nippon Telegraph and Telephone Public Corporation, Musashino-shi, Tokyo, Japan.

II. Spoken Word Recognition System

A block diagram of the spoken word recognition system used in this investigation is shown in Fig. 1 [1]. Reference patterns are stored as phoneme templates. A word dictionary containing symbolic phoneme spellings is also included. Some phonemes have two or more kinds of templates in order to cope with the variation of speech spectra due to coarticulation. In this investigation two sets of Japanese words are used in the recognition test. One is the set of ten digits and the other is 67 airport names. Table I gives these lists.

Input speech is passed through a low-pass filter at 3.4 kHz, sampled at an 8 kHz rate, and divided into a succession of 15 ms duration segments. 0th through Nth order correlation coefficients are extracted from each segment. N is set to 10 in this experiment. A spectral similarity calculation between the jth segment of input speech and each reference phoneme template is carried out by a logarithmic likelihood measure $l_j(x_i)$, where x_i means the ith phoneme. The logarithmic likelihood $l_j(x_i)$ can be measured using (1) with the correlation coefficients of input speech $\zeta_\tau(j)$ and linear prediction coefficients of the phoneme template $a_\tau(x_i)$, where τ is the time delay. The phoneme templates are stored in the form of the correlation function of the linear prediction coefficients (LPC's) $A_\tau(x_i)$.

$$l_j(x_i) = -\log\left\{\sum_{\tau=-N}^{N} A_\tau(x_i)\,\zeta_\tau(j)\right\}$$

$$A_\tau(x_i) = \sum_{k=0}^{N-\tau} a_k(x_i)\,a_{k+\tau}(x_i). \tag{1}$$

Word identification is based on the total likelihood obtained by summing up the logarithmic likelihood over the input word-length using a likelihood matrix in which each row corresponds to an input segment and each column corresponds to a phoneme template, as shown in Fig. 1. In each comparison with candidate words, input speech is segmented by phoneme boundaries, and one or more 15 ms duration segments are associated with each phoneme according to the phoneme spell-

0096-3518/80/0400-0129$00.75

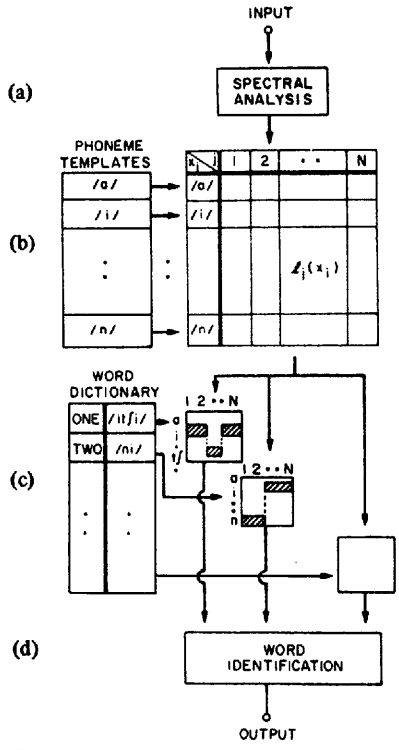

Fig. 1. Block diagram of the spoken word recognition system. (a) Extraction of autocorrelation coefficients for the input speech wave. (b) Computation of log likelihood matrix in which row and column correspond to phoneme and segment number of the input speech, respectively, and the element denotes the degree of log likelihood between the reference template and the input speech segment. (c) Computation of total likelihood between each candidate word and the input speech. (d) Selection of the word which has the maximum total likelihood.

ings in the word dictionary. As only the order and the typical length of the duration of each phoneme are written in the word dictionary, there are many possibilities of segmentation into phoneme groups. The optimum segmentation can be done automatically by a dynamic programming procedure [9] on the likelihood matrix. This procedure is carried out so that the total value of the logarithmic likelihood is maximized over all the segments of input speech for all possible segmentations. The total likelihood is used as the similarity of the input speech sample to the candidate word. The input speech sample is designated as that word whose similarity is larger than that of any other candidate.

Phoneme templates are extracted from training words uttered by each speaker. The phoneme boundaries of these training speech samples are automatically determined and the spectra of each phoneme are averaged in the correlation coefficient domain to make a phoneme template by the following method. At first, the phoneme boundaries are determined using the phoneme template of a previous speaker by the same procedure described above for word identification. In this case, since the training word is known, the procedure is much simpler than that used in the identification. An initial set of phoneme templates is established based on these boundaries. The process continues by determining phoneme boundaries all over again using this initial set of templates obtained from the speaker. This process is iterated with updated templates until the phoneme boundaries become stable.

TABLE I
WORD LISTS: (a) 10 DIGITS. (b) 67 AIRPORT NAMES.

ichi	/it/i/	(one)	roku	/roku/	(six)
ni	/ni/	(two)	nana	/nana/	(seven)
san	/san/	(three)	hachi	/hat/i/	(eight)
yon	/jon/	(four)	kyu	/kju: /	(nine)
go	/go: /	(five)	ju	/dju: /	(ten)

(a)

Sapporo	Yonago	Okinoerabu	Mombetsu
Asahikawa	Izumo	Yamagata	Nakashibetsu
Memambetsu	Tokushima	Sendai	Okinawa
Wakkanai	Takamatsu	Hachijojima	Kumejima
Kushiro	Kochi	Oshima	Minamidaito
Obihiro	Hiroshima	Miyakejima	Miyako
Hakodate	Ube	Toyama	Tarama
Akita	Matsuyama	Kanazawa	Ishigaki
Aomori	Oita	Fukui	Yonakuni
Hachinohe	Fukuoka	Tottori	Chitose
Hanamaki	Miyazaki	Nagoya	Komatsu
Tokyo	Kagoshima	Kitakyushu	Omura
Osaka	Tanegashima	Nagasaki	Naha
Nankishirahama	Yakushima	Kumamoto	Misawa
Niigata	Kikaijima	Fukue	Rishiri
Okayama	Amamioshima	Sado	Okushiri
Oki	Tokunoshima	Iki	

(b)

III. TRAINING BY A FRACTION OF THE ENTIRE VOCABULARY

In this paper a new training procedure which uses the utterances of only a fraction of the entire vocabulary as training speech samples is proposed and tested. In this method all the phoneme templates are not necessarily extracted from the training samples. The spectra of lacking phonemes are optimally estimated from other extracted phonemes' spectra. Also, since the spectra of the phonemes directly extracted from training samples may be influenced by coarticulation peculiar to the training samples, they are modified and adjusted to fit the spectra of nontraining words.

A. Pretraining Procedure–Multiple Regression Analysis

Prior to training, some statistical analyses of the phoneme spectra are made. Several speakers (pretraining speakers) are required to utter the entire vocabulary once. These speakers are completely distinct from the speakers who will speak training and test samples. The speech wave of each sample is segmented by phoneme boundaries automatically by the method described in Section II and two sets of phoneme templates are obtained by averaging the correlation coefficients. One set is extracted by averaging over the entire vocabulary. The other set is obtained by averaging only over those words designated as training words. These two sets of averaged correlation coefficients are transformed into log-area-ratios (LAR's) [2]. The LAR parameters can be obtained by arctanh transformation of PARCOR coefficients.

Multiple regression analysis [8] is applied to obtain a relation between the set obtained over the whole vocabulary and the set obtained over the fraction of the vocabulary. Let u_{im} be the mth LAR of phoneme i extracted from the whole vocabulary, and v_{jl} be the lth LAR of phoneme j extracted from the fraction of the vocabulary. It is assumed that there is a relation of the form shown in (2) between $U_i = (u_{i1}, u_{i2}, \cdots,$

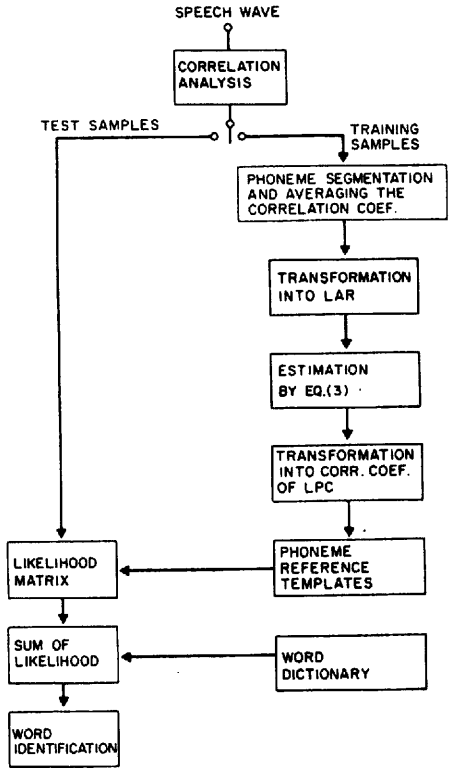

Fig. 2. Block diagram of the fractional vocabulary training and recognition procedures.

$u_{iN})'$ and $V_j = (v_{j1}, v_{j2}, \cdots, v_{jN})'$ irrespective of speakers [4].

$$U_i = \Phi_{ij} V_j + E_{ij} \tag{2}$$

where Φ_{ij} is an $N \times N$ matrix and E_{ij} is a residual. The author's previous investigation [7] shows the possibility of this assumption on the modeling of speech individuality. LAR parameters are appropriate to this transformation process because of their stability characteristics and nearly uniform spectral sensitivity [3].

Φ_{ij} is estimated by the multiple regression analysis method using the speech data of all the pretraining speakers. At the same time, the multiple correlation coefficient R_{ijl} between the lth LAR of phoneme i, estimated from phoneme j by the model (2) and actual value of the lth LAR of phoneme i, is calculated. Matrix Φ_{ij} is calculated for every combination of i and j, and matrix $\{R_{ijl}\}$ is calculated for every combination of i, j, and l. For each phoneme i, LAR vectors U_i of all the pretraining speakers are averaged to produce a vector $\bar{x}_i$.

The matrix Φ_{ij}, the set of multiple correlation coefficients $\{R_{ijl}\}$, and the vector $\bar{x}_i$ will be used, as described in the following section, to provide phoneme estimates.

B. Training Procedure

Every new speaker is required to utter a set of training samples which is a fraction of the whole recognition vocabulary. Fig. 2 shows the procedure of training and recognition. First, the speech wave of each training sample is converted to a time sequence of correlation coefficients. The sequence is segmented by phoneme boundaries automatically, and correlation coefficients of all the segments of each phoneme are averaged over all the training samples. The averaged correlation coefficients are transformed into LAR's. Then, LAR parameters of all phonemes, whether or not they have been extracted from the training samples, are estimated based on the LAR parameters of the phonemes that have been extracted directly from the training samples. The following equation based on (2) is used for phoneme estimation:

$$\hat{x}_{ik} = \left(r \Big/ \sum_{j \in x_L} w_{ij} \right) \sum_{j \in x_L} w_{ij} \Phi_{ij} x_{jk} + (1 - r)\, y_{ik}: \; i \in X \tag{3}$$

where $\hat{x}_{ik}$ is an estimated N-dimensional LAR vector of phoneme i and x_{jk} is an N-dimensional LAR vector of phoneme j extracted from training samples. k is the training speaker number, X is a set of all the phonemes, and X_L is a subset of X which can be extracted from training samples. Equation (3) consists of two terms. The first term relates the entire set of phoneme templates to the training set templates by means of the transformation rules Φ_{ij}, which is the $N \times N$ matrix estimated in the pretraining. Included in this term is a set of weighting coefficients w_{ij} defined as follows:

$$w_{ij} = (1/N) \sum_{l=1}^{N} R_{ijl}^2 \tag{4}$$

where R_{ijl} is the multiple correlation coefficient calculated at the same time as Φ_{ij}. The second term in (3) relates the entire set of phoneme templates directly to either the training set templates or the averaged pretraining templates for those templates not included in the training set. The weighting coefficient r, set between 0 and 1, determines the relative mixture of the direct and indirect terms in the estimation of $\hat{x}_{ik}$. y_{ik} is defined as follows:

$$y_{ik} = \begin{cases} x_{ik}: & i \in X_L \\ x_{i'k}: & i \notin X_L, \; i' \in X_L \\ \bar{x}_i: & i \notin X_L, \; i' \notin X_L \end{cases} \tag{5}$$

where i' means another reference pattern of phoneme i that is prepared to treat the effect of the coarticulation, and $\bar{x}_i$ is the mean LAR vector of the phoneme i which is extracted in the pretraining step.

The estimated LAR parameters are transformed into correlation coefficients of linear prediction coefficients and stored in the reference phoneme pattern area.

After training, each input speech wave is analyzed and transformed into a time sequence of correlation coefficients. Using these correlation coefficients and the correlation coefficients of linear prediction coefficients of every phoneme, a likelihood matrix is calculated. Using the word dictionary, the amounts of logarithmic likelihood are summed over the wordlength using a dynamic programming procedure, and word identification is accomplished based on this total likelihood value.

IV. Recognition Experiment Using Spoken Digits

A. Experimental Conditions

The first evaluation to test the effectiveness of the training method used a vocabulary consisting of Japanese digits. Ten

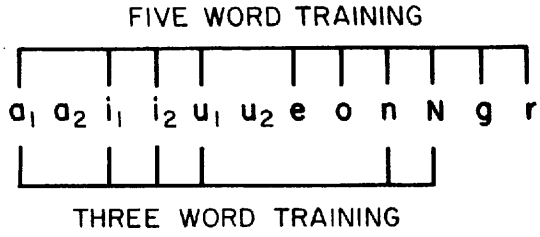

Fig. 3. Phoneme reference templates for the spoken digit vocabulary indicating those templates which are included in the fractional vocabulary training sets.

digits were uttered by 30 male speakers and 30 female speakers. Each speaker uttered these words in random order and each word was uttered 11 times. Sets of three or five digits were chosen as training words. The three-digit set consisted of 2, 3, and 9, while the five-digit set contained 0, 2, 3, 5, and 9. These words were selected for the reason that their recognition accuracies were relatively low for the condition in which there was no training, that is, when common reference patterns were used for all speakers.

From the 11 samples of each digit, the first sample is used for training. The latter ten utterances are used for recognition tests. The number of test samples is 100 for each speaker.

All the phonemes of the reference templates and those extracted from training samples are presented in Fig. 3. There are two kinds of reference patterns for the phonemes /a/, /i/, and /u/ in order to compensate for coarticulation. Since unvoiced consonants, for example, the /tʃ/ sound in /itʃi/ and the /s/ sound in /san/, have relatively short duration times and do not play important roles in digit recognition, they are ignored in this experiment. Out of the total of 12 phonemes, ten and six phonemes are extracted from the five-word and three-word training sets, respectively.

Φ_{ij}, w_{ij}, and $\bar{x}_i$, are obtained during the pretraining procedure using the utterances of a population of 29 speakers independent of the test speakers. These values are calculated separately for male and female speakers.

The recognition results are compared with the results of three additional experiments. In the first experiment, training is omitted and $\bar{x}_i$, which is the template averaged over the 29 pretraining speakers, is used as $\hat{x}_{ik}$ in (3) for all test speakers. In the second one, the modification of the x_{ik} directly extracted from the training set is not done, and y_{ik} is used as $\hat{x}_{ik}$. In the last experiment, all of the ten digits are used for training. Summarizing, the following four experiments distinguished by training conditions are carried out.

1) Fractional vocabulary training with estimation–$\hat{x}_{ik}$ is specified by (3).

2) Fractional vocabulary training without estimation–$\hat{x}_{ik} = y_{ik}$.

3) No training–$\hat{x}_{ik} = \bar{x}_i$.

4) Whole vocabulary training.

B. General Performance

Fig. 4 shows the mean error rate averaged over 30 speakers, and ten digits on each training condition with the weight $r = 0.5$. Fig. 5 shows the relation between the value of r and the mean

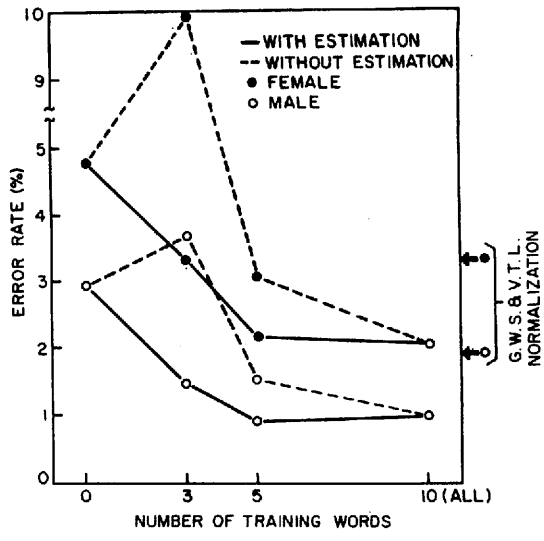

Fig. 4. Mean recognition error rate for spoken digit vocabulary for each condition compared with the previous results obtained after (GWS) glottal wave spectrum and (VTL) vocal tract length normalization [5].

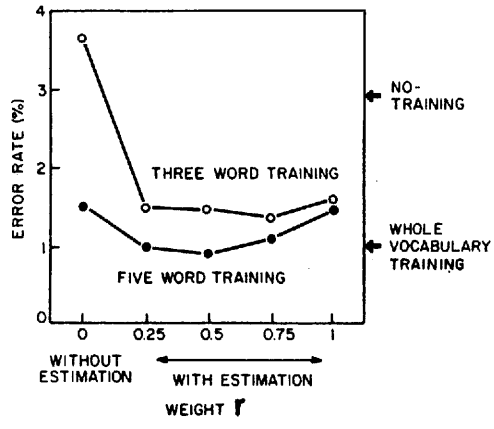

Fig. 5. Relation between weight r and recognition error rate (male speaker).

error rate which is obtained with the fractional vocabulary training method with estimation.

It can be seen in Fig. 4 that the fractional vocabulary training method with estimation using five words, which is half of the ten-digit vocabulary, produced error rates of 0.9 percent and 2.1 percent for male and female speakers, respectively. These values are almost the same as the error rates obtained with the whole vocabulary training procedure. Training using only three words produced error rates of 1.4 percent and 3.3 percent for male and female speakers, respectively, which are about 1.5 percent lower than those obtained without training. When estimation is not used in the fractional vocabulary training method and the number of training samples is so small that only a part of the phonemes of the reference patterns are extracted, the error rate becomes very large. The training method using three words brings slightly better results than the results obtained by the same author [5] when the effects of glottal spectrum and vocal tract length are approximately normalized.

From the results shown in Fig. 5, it can be concluded that a good value of the weight r is about 0.5. The error rate in-

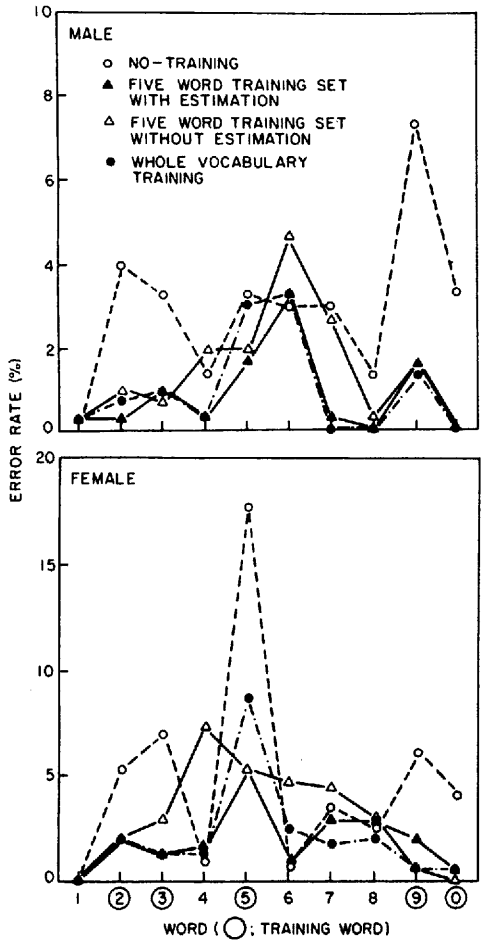

Fig. 6. Recognition error rate of each word for spoken digit vocabulary (five-word training set).

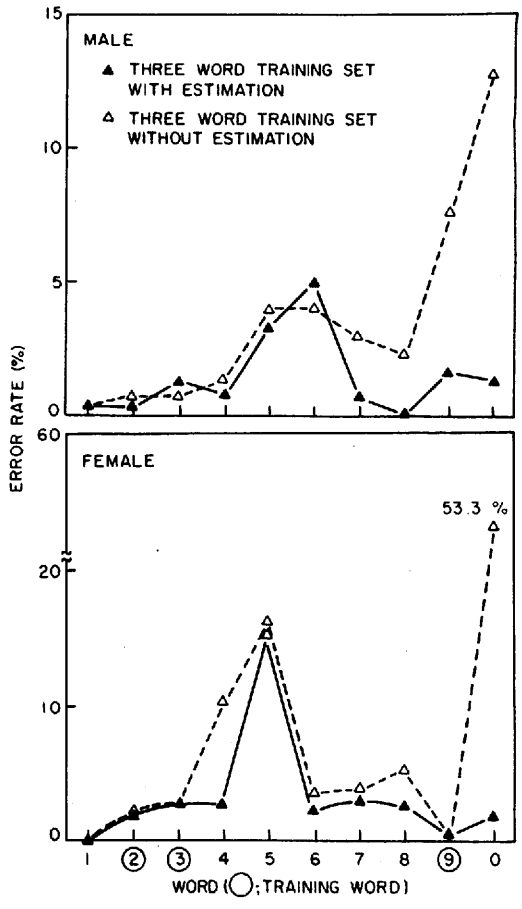

Fig. 7. Recognition error rate of each word for spoken digit vocabulary (three-word training set).

creases not only when $r = 0$, but also when $r = 1$, the latter condition indicating that the term y_{ik} is not used.

The results of these recognition experiments using spoken digits show the effectiveness of the training method which uses only a fraction of the whole vocabulary.

C. Discussion of the Results of Spoken Digit Recognition

Fig. 6 shows the results for each word of the five word training condition compared with no training and whole vocabulary training for both male and female speakers. It can be seen that the error rates for the words 4, 6, and 7, which are not included in the training set, become high for the condition of training without estimation. But when estimation is used, the rates become very low and the distribution of the error rates among ten digits becomes similar to that obtained for the whole vocabulary training condition. The results for female speakers are quite similar to those for male speakers.

From the fact that all phonemes but /a_2/ and /u_2/ are extracted from the five-word training set (see Fig. 3), the effects of the estimation indicated in Fig. 6 can be regarded as equivalent to the removal of the influence of coarticulation which is peculiar to the training set.

Fig. 7 shows the results using the three-word training set for male and female speakers. The estimation process greatly reduces the error rate for the digits of 7, 8, and 0, none of which are included in the training set. The error rate for 9, which is included in the training set, is also reduced by estimation for male speakers. Most of the recognition confusion for the training without the estimation condition is between 9 (/kju:/) and 0 (/rei/), and 0 and 2 (/ni/). It seems likely that the improvement using estimation mainly results from the estimation of the spectral distinctiveness of the phoneme /e/.

The weight w_{ij} used in the estimation formula is the correlation between the true LAR value of phoneme i and that estimated from phoneme j. Fig. 8 shows actual values of w_{ij}, used for the five-word training set, averaged over all i. The weight values between identical phonemes (w_{ij}; $j = i$ or $j = i'$) and those between different phonemes are averaged separately. The weight between identical phonemes is 1.3 to 1.6 times as large as the weight between different phonemes. Similar results are obtained for the three-word training set condition.

V. Recognition Experiment Using a 67-Word Vocabulary

A. Experimental Conditions

The training method tested in the previous section was also applied to a relatively large vocabulary word recognition system. The vocabulary of this system is 67 Japanese words, which are names of airport locations in Japan. Speech samples uttered by 30 male speakers are used in the experiment. The 67 words are uttered sequentially and this sequence is repeated

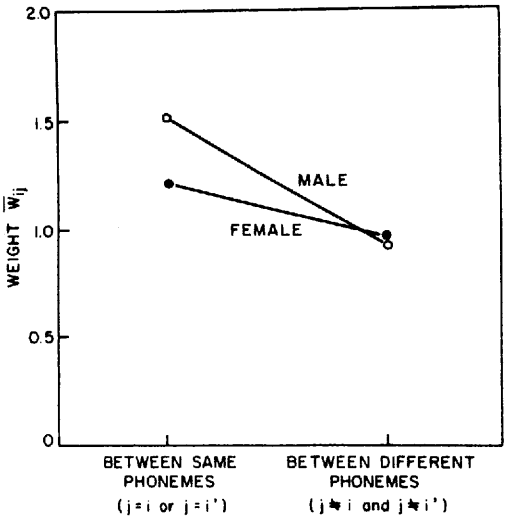

Fig. 8. Averaged weight value $\bar{w}_{ij}$ shown separately for pairs of the same phonemes and pairs of different phonemes.

TABLE II
PHONEME REFERENCE TEMPLATES FOR THE 67-WORD VOCABULARY

Class	Number	Reference Templates
Vowels	9	a i_1 i_2 i_3 u_1 u_2 u_3 e o Δ Δ Δ Δ Δ Δ Δ Δ Δ
Transition between Vowels	6	ia io iu ai oi ui Δ Δ Δ Δ Δ Δ
Voiced Consonants	13	b d g z_1 z_2 r_1 r_2 w N m n_1 n_2 ŋ ¶ Δ ¶ ¶ ¶ Δ ¶ Δ Δ Δ Δ Δ ¶
Unvoiced Consonants	13	p t ts tʃ k_a k_1 k_u k_e k_o h_1 h_2 f s Δ Δ Δ Δ Δ Δ Δ ¶ Δ ¶ Δ Δ Δ
Pause (Noise)	1	* Δ
Total	42	

¶—Phoneme which is extracted from the 25-word training set, but not extracted from the 12-word training set.
Δ—Phoneme which is extracted both from the 25-word training set and the 12-word training set.

five times. All or a part of the speech samples of the first sequence are used for training and the latter four sequences are used for recognition.

From the 67 words, 12 or 25 words are selected and used for fractional vocabulary training. These words are selected by the same principle used in the spoken digit recognition experiment.

The number of phoneme reference templates used in this system is 42, including a pause/noise spectrum. These phonemes are presented in Table II. The phonemes which can be extracted from the fractional vocabulary words are marked in this figure. The number of phonemes extracted from the 25-word training set is the same as that extracted from the whole vocabulary training set. From the 12-word training set, eight phonemes cannot be extracted.

Similar to the digit recognition system, there are several phonemes which have more than one reference pattern. But in this experiment, all these reference patterns are treated independently, and the expression (5) is changed as follows:

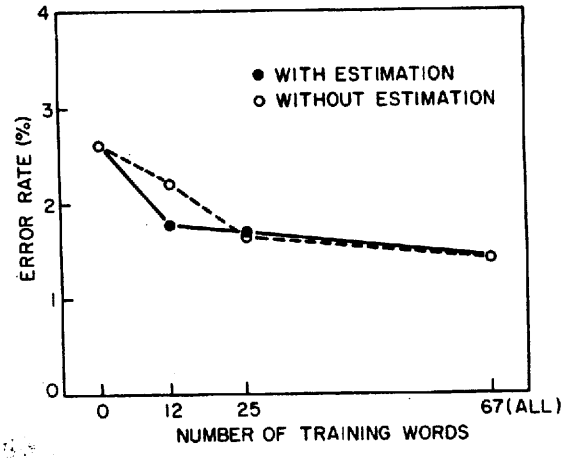

Fig. 9. Mean recognition error rate for the 67-word vocabulary for each condition.

$$y_{ik} = \begin{cases} x_{ik}: & i \in X_L \\ \bar{x}_i: & i \notin X_L. \end{cases} \tag{6}$$

The matrix Φ_{ij}, weight w_{ij}, and vector $\bar{x}_i$ which are used in the training procedure are calculated using the speech samples of 25 pretraining speakers distinct from the speakers who provide the training and test samples. The weight r is fixed at 0.5.

B. General Performance

Mean error rate averaged over 30 speakers and 67 words for each condition is shown in Fig. 9. When the 25-word set is used for training, the estimation procedure does not improve the error rate. However, when the 12-word set is used for training, the error rate can be reduced by the estimation procedure from 2.2 percent to 1.8 percent. The 12-word set training is much easier than whole vocabulary training and provides a much lower error rate compared with no training. This result shows that the new method using a fractional vocabulary training set is very useful and practical.

C. Discussion of the Results of the 67-Word Vocabulary Recognition System

Error rates for training words and nontraining words are averaged separately and presented in Fig. 10 for the conditions of no training, the 12-word training set, and the whole vocabulary training set. Error rates averaged over all words are also shown in this figure. When the 12-word training set without estimation is used, the mean error rate averaged over the 12 training words becomes much smaller than that obtained for the no training condition, although the averaged error rate for 55 nontraining words is slightly increased. The estimation process makes the mean error rate of nontraining words low; conversely, the mean error rate of training words becomes somewhat high. But the mean error rate averaged over all words becomes small since the number of nontraining words is much greater than the number of training words. When the estimation procedure is carried out, the error rates of training and nontraining words are nearly equal to the error rates which are obtained for the whole vocabulary training condition.

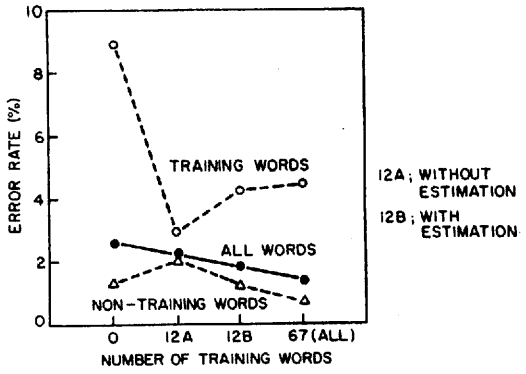

Fig. 10. Recognition error rates for training words and nontraining words for the 67-word vocabulary.

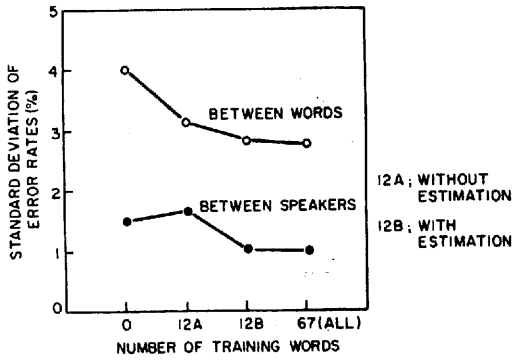

Fig. 11. Standard deviation of error rates among words and among speakers.

In order to examine the distribution of error rates among words and speakers, the standard deviations of the error rates are calculated and shown in Fig. 11 for every training condition. With no training, the distributions of the error rates both among words and among speakers are very wide. For the 12-word training condition with estimation, both distributions become as narrow as in the case of the whole vocabulary training condition.

VI. Conclusion

In order to obtain high accuracy in a spoken word recognition system with a large vocabulary irrespective of speakers, a phoneme-based, quasi-speaker independent system which includes a new training method using a small training sample from each speaker is proposed and tested by a ten-digit recognition system and 67-word recognition system. In this method each new speaker is required to provide utterances of only a fraction of the entire vocabulary as a training set. Speech samples from the training set are converted into time sequences of correlation coefficients and segmented by phoneme boundaries automatically. For each phoneme, mean correlation coefficients over all segments of the training set are calculated and transformed into log-area-ratio parameters. A set of transformation rules is used to estimate phoneme templates for the entire vocabulary from phoneme templates included in the training set. The transformation rules are obtained in a pretraining procedure in which a group of speakers provides utterances of the entire vocabulary. Multiple regression analysis is used to obtain linear estimates of the entire phoneme templates set in terms of the set designated as training templates. This group of speakers is generally distinct from the group of training speakers. Thus, since the transformation rules are established independent of the training speakers, the entire procedure can be considered a quasi-speaker-independent system, or a hybrid speaker-dependent/speaker-independent system.

Results of the ten-digit recognition system using 30 male and 30 female voices indicate that a five-word training set is sufficient to obtain the same recognition accuracy that can be obtained using whole vocabulary as a training set. A three-word training set is quite effective to largely reduce the error rate compared with no training. The mean recognition accuracy with the five-word training procedure is 99.1 percent and 97.9 percent for male and female speakers, respectively. On the other hand, the mean recognition accuracy with the three-word training procedure is 98.6 percent and 96.7 percent, respectively. In the 67-word recognition experiment, it has been demonstrated that a 12-word training set is highly effective to improve the performance. The mean recognition accuracy of 98.2 percent was obtained for male speakers with a 12-word training set. This shows the applicability of the new training method to actual large vocabulary recognition systems.

A special characteristic of this method is that, not only reference templates not included in the training samples are estimated from those that are, but the phoneme templates which are extracted from training samples are automatically modified and adjusted to the whole vocabulary including nontraining words. From the principle of this training procedure it can be suggested that training samples may be completely independent

of the recognition vocabulary and perhaps could be some special short sentence.

Four areas need further investigation:

1) selection of words or a sentence for training,

2) optimization of the weight coefficients in the estimation equation,

3) an applicability test of this method to other large vocabulary word recognition systems or connected word recognition systems, and

4) improvement of the method by the combination of a supervised [6] or nonsupervised [7] training procedure which is performed at the same time as recognition.

Acknowledgment

The author especially wishes to acknowledge the guidance provided by Dr. S. Saito, Director of the Saito Research Section, the advice and the offer of speech data and some programs from M. Kohda, Staff Engineer of the 4th Research Section, and H. Nagashima, Engineer of the Saito Research Section, Musashino Electrical Communication Laboratory. The author also wishes to thank Dr. A. E. Rosenberg at Bell Laboratories for the revision of this paper.

References

[1] S. Saito and M. Kohda, "Spoken word recognition using the restricted number of learning samples," in *Proc. 1976 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1976, sect. 6.11.

[2] S. Furui and F. Itakura, "Talker recognition by statistical features of speech sounds," *Trans. IECE Jap.*, vol. 56-A, p. 717, Nov. 1973.

[3] R. Viswanathan and J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, p. 309, June 1975.

[4] S. Furui, "Efficient learning method of talker differences in spoken word recognition," in *Proc. Fall Meeting Acoust. Soc. Jap.*, Oct. 1977, no. 3-1-19.

[5] —, "Analysis of phonemic and personal information in speech spectral pattern," in *Proc. Nat. Conv. Rec. IECE Jap.*, Mar. 1976, no. S11-10.

[6] B. T. Lowerre, "Dynamic speaker adaptation in the Harpy speech recognition system," in *Proc. 1977 IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1977, no. 23.2.

[7] S. Furui, "On the separation of personal and phonemic information in the speech wave," in *Proc. Fall Meeting Acoust. Soc. Jap.*, Oct. 1974, no. 2-2-10.

[8] N. L. Johnson and F. C. Leone, "Statistics and experimental design in engineering and applied sciences," vol. 1, 2nd ed. New York: Wiley, 1977, p. 464.

[9] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proc. Int. Conf. Acoust.*, Budapest, Hungary, 1971, paper 20C-13.

Sadaoki Furui (M'79) was born in Tokyo, Japan, on September 9, 1945. He received the B.S., M.S., and Ph.D. degrees in mathematical engineering and instrumentation physics from Tokyo University, Tokyo, Japan, in 1968, 1970, and 1978, respectively.

After joining the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation, in 1970, he studied the analysis of speaker characterizing information in the speech wave, and its application to speaker recognition and interspeaker normalization in speech recognition. He is currently a Staff Engineer at Musashino Electrical Communication Laboratory. From December 1978 to December 1979 he joined the staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, NJ, as an Exchange Visitor working on speaker verification.

Dr. Furui is a member of the Acoustical Society of Japan and the Institute of Electronics and Communication Engineers of Japan.