

論文 / 著書情報
Article / Book Information

Title	Unsupervised Speaker Adaptaion Based on Hierarchical Spectral Clustering
Author	SADAOKI FURUI
Journal/Book name	IEEE Trans. on ASSP, Vol. 37, No. 12, pp. 1923-1930
発行日 / Issue date	1989, 12
権利情報 / Copyright	(c)1989 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.

Unsupervised Speaker Adaptation Based on Hierarchical Spectral Clustering

SADAOKI FURUI, SENIOR MEMBER, IEEE

Abstract—This paper proposes a new automatic speaker adaptation algorithm for speech recognition, in which a small amount of training material of unspecified text can be used. This algorithm is easily applicable to VQ-based speech recognition systems consisting of a VQ codebook and a word dictionary, in which each word is represented as a sequence of codebook entries. In the adaptation algorithm, the VQ codebook is modified for each new speaker, whereas the word dictionary is universally used for all speakers. The important point of this algorithm is that a set of spectra in training frames and the codebook entries are clustered hierarchically. Based on the deviation vectors between centroids of the training frame clusters and the corresponding codebook clusters, adaptation is performed hierarchically from small to large numbers of clusters. The spectral resolution of the adaptation process is therefore improved accordingly. Results of recognition experiments using utterances of 100 Japanese city names show that the adaptation reduces the mean word recognition error rate from 4.9 to 2.9 percent. Since the error rate for speaker-dependent recognition is 2.2 percent, the proposed adaptation method is highly effective.

I. INTRODUCTION

EXISTING methods for achieving highly accurate speaker-independent speech recognition can be classified as follows:

- 1) combining new feature parameters which have relatively small interspeaker variability, such as dynamic spectral features [1], with conventional instantaneous spectral parameters;
- 2) designing multiple reference templates or statistical models to cover the range of individual variation [2], [3]; and
- 3) automatically adapting reference templates to each new speaker or normalizing (reducing) interspeaker variation in each input speech, based on the transformation rules obtained using a few training words or a short sentence [4]–[6].

Although the first two methods seem easier and more straightforward than the third, and their effectiveness has been experimentally confirmed for various recognition tasks, these methods have two problems. One is that actual performance is limited since the spectral features of different phonemes uttered by different speakers sometimes overlap [7]. The other is that these methods tend to increase the computation of recognition systems. Even when these methods are used, recognition systems gen-

erally still require a retraining mechanism in practical applications to cope with variations due to different telephone sets or microphones, different background noise conditions and room acoustics, variable bandwidth limitations of different speech transmission systems, and so on.

Therefore, it is crucial to establish the third method, the automatic adaptation mechanism, and possibly combine it with the first two methods. Humans have been found to possess a similar adaptation mechanism. Although the first several words uttered by a speaker new to the listener may be unintelligible, the listener quickly becomes accustomed to the new voice, and intelligibility increases after hearing several words [8].

The following six methods have been investigated for realizing the speaker adaptation mechanism.

- 1) Vocal cord spectrum and vocal tract length are estimated for each speaker, and individual variations of these values are normalized. They are respectively extracted from the averaged overall spectrum, that is, the overall pattern of the long-time averaged spectrum, and from the linear expansion or contraction coefficient for the spectrum along the frequency axis [9].
- 2) Phoneme reference templates adapted to each speaker are generated from the templates extracted from training words using a set of transformation rules. These rules indicate the relationship between the phoneme templates for the entire vocabulary and those extracted from a small number of training words, based on statistical modeling [4].
- 3) Phonetic (linguistic) information and individual information in speech spectra are automatically separated according to statistical modeling of spectral structure [10].
- 4) Many sets of phoneme templates representing speaker variability are stored beforehand, and the most suitable set for the new speaker is automatically selected [10], [11].
- 5) In VQ-based word recognition, the VQ codebook for a reference speaker is replaced by that for a new speaker based on the one-to-one correspondence (mapping) rule of vectors. The correspondence rule is automatically determined by DP-matching between training and reference utterances, both of which are represented by the sequence of codebook elements [5].
- 6) In HMM-based recognition, HMM speech models derived from one "prototype" speaker are transformed so

Manuscript received May 24, 1988; revised February 18, 1989.

The author is with NTT Human Interface Laboratories, 3-9-11 Midoricho, Musashino-shi, Tokyo, 180 Japan.

IEEE Log Number 8931353.

that they model the speech of a new speaker. This transformation is accomplished by probabilistic spectral mapping from one speaker's spectral space to another's [6].

Automatic training or adaptation methods are generally classified into supervised (text-dependent) methods in which training words or sentences are known, and unsupervised (text-independent) methods in which arbitrary utterances can be used. Of the six methods, 2), 5), and 6) are supervised. Even in unsupervised methods, phonemes have usually been explicitly determined by speech recognition processes embedded in the training process, and these results are used for adaptation. A completely unsupervised method should have a wide range of applications. Human adaptation performance was also found to be independent of the content of training utterances [8].

Shiraki *et al.* [12] recently proposed a new effective unsupervised method for adapting the codebook for a segment vocoder to a new speaker. In this method, spectral segments are hierarchically clustered and adapted, while maintaining continuity between adjacent clusters. Although correspondence between each codebook element and linguistic information need not be maintained during adaptation in speech coding, it is crucial that it be maintained in the training for speech recognition using a limited number of training words. This is necessary because, if the correspondence is not maintained, the sequences of codebook entries (code words) which represent each vocabulary word cannot be produced without uttering every word. In other words, if only a fraction of the vocabulary words or a short sentence which does not include any vocabulary word is used for training, the code-word sequences must be maintained and correspondence between each codebook element and its linguistic information must be maintained in the training procedure. Therefore, the speaker-adaptation method proposed for speech coding may not necessarily be applicable to speech recognition.

This paper investigates the feasibility of the speaker-adaptation method for the segment vocoder to speech recognition, and examines various important features of this method. Section II of this paper describes the adaptation method and its implementation in a VQ-based word recognition system. Section III presents the experimental results, and Section IV presents supplementary experiments. Section V summarizes the paper.

II. UNSUPERVISED SPEAKER-ADAPTATION METHOD

A. Recognition System Structure Including Adaptation

Fig. 1 is a block diagram of the VQ-based spoken word recognition system, SPLIT (strings of phoneme-like templates method [13]), including the speaker-adaptation process investigated in this paper. In the SPLIT system, each word is represented by a set of multiple sequences (word dictionary) of VQ-codes (pseudophoneme entries). The VQ codebook is adapted to each speaker, whereas the sequences of VQ-codes for each word entry in the dictionary are maintained for all speakers. Individual variations on how a word is uttered are modeled by multiple code sequences. The VQ codebook adaptation proposed here

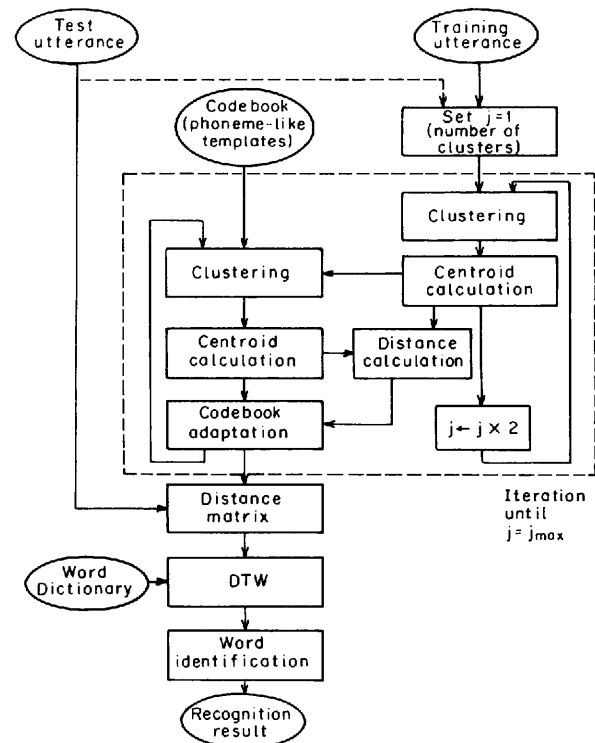


Fig. 1. Principal structure of a word recognition system with the speaker adaptation mechanism.

can be performed using any word set or short sentence. The adaptation can be done using even unknown input utterances to be recognized.

The most important point of this adaptation method is in the use of hierarchical clustering of both spectra in the training speech samples and those in the reference VQ codebook. Based on hierarchical clustering, adaptation is performed step by step from global to local individuality using the deviation vectors between centroids of corresponding clusters. This hierarchical algorithm is highly effective for obtaining good adaptation performance.

Input utterances are then recognized by the SPLIT method using the adapted codebook elements. The adaptation algorithm proposed by Shiraki *et al.* [12] is described precisely in the following two subsections, B and C.

B. Adaptation Procedure

First, an initial VQ codebook and a VQ-indexed word dictionary are prepared. The initial codebook, which we call universal codebook, is produced by clustering the voices of multiple speakers, and is commonly used as the initial condition for every new speaker. The following adaptation procedures are performed for the VQ codebook elements. The size of the VQ codebook is maintained throughout the adaptation process.

1) The initial and the maximum number of clusters are set at 1 and $j_{\max}$, respectively.

2) The deviation vector between the centroid of all training utterance spectra of a new speaker and the centroid of all the universal codebook elements is calculated.

3) All of the universal codebook elements are shifted by this deviation vector so that the corresponding centroids coincide.

4) The number of clusters (j) is doubled, and the training utterance spectra are clustered by the LBG algorithm [14].

5) The centroid of each training utterance cluster is calculated, and the distances between all the combinations of codebook elements and training utterance cluster centroids are measured. Codebook elements are then clustered so that each element belongs to its nearest training utterance cluster.

6) The centroid for each codebook element cluster is calculated, and the deviation vectors between training utterance cluster centroids and corresponding codebook element cluster centroids are produced.

7) Using these deviation vectors, codebook elements are shifted, while continuity between adjacent clusters is maintained. (The precise procedure for maintaining continuity is explained in Subsection II-C.)

8) If $j = j_{\max}$, stop. If $j < j_{\max}$, return to 4).

Codebook elements are thus hierarchically adapted to a new speaker by repeating steps 4)–8).

C. Codebook Element Shift Maintaining Continuity Between Adjacent Clusters

As described in the previous subsection, codebook elements are shifted using deviation vectors between centroids of training speech clusters and those of corresponding codebook element clusters while maintaining continuity between adjacent clusters. This is accomplished by considering the deviation vectors of not only the cluster in which the codebook element is included, but also other clusters when determining the shifting vectors.

When the centroid of the m th codebook element cluster is represented by u_m , and that of corresponding training speech cluster by v_m , the deviation vector between these two centroids is

$$p_m = v_m - u_m \quad (m = 1, \dots, j), \quad (1)$$

where j is the total number of clusters. For each codebook element, c_i , the deviation vectors for all clusters, $\{p_m\}$, are summed with weighting factors, $\{w_{im}\}$, to produce the shift vector, Δ_i :

$$\Delta_i = \left(\sum_{m=1}^j w_{im} p_m \right) / \left(\sum_{m=1}^j w_{im} \right). \quad (2)$$

The weighting factor w_{im} is the inverse of the distance between c_i and u_m to the power of α :

$$w_{im} = 1 / \|c_i - u_m\|^\alpha. \quad (3)$$

The codebook element, c_i , is then shifted to c'_i as

$$c'_i = c_i + \Delta_i. \quad (4)$$

Fig. 2 illustrates the hierarchical adaptation procedures from the beginning to the four-cluster stage. This adaptation method is hereafter referred to as Method I.

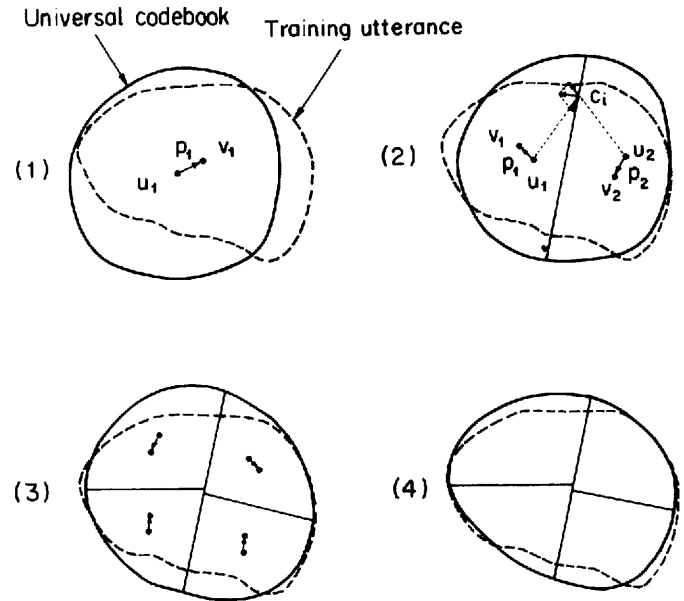


Fig. 2. Hierarchical codebook adaptation algorithm maintaining continuity between adjacent clusters.

To evaluate the effectiveness of this method, it is compared to an alternative method which we call Method II. In Method II, codebook elements are directly adapted using a fixed number of $j_{\max}$ clusters without hierarchical adaptation. In other words, the adaptation steps from 4) to 7) in Subsection II-B are performed under the condition of $j = j_{\max}$ without iteration.

III. RECOGNITION EXPERIMENTS

A. Recognition Task and Feature Extraction

Japanese 100-city name utterances have been used to evaluate the adaptation algorithms described in the previous section, in isolated word, speech recognition experiments. Speech signals are sampled at 8 kHz, and spectrally represented by the sequence of the first 10 LPC cepstral coefficients and logarithmic energy at every 16 ms frame interval. The time derivatives of cepstral coefficient and logarithmic energy sequences are approximated respectively by first-order regression coefficients over a finite length window of 88 ms, centered around the current frame.

Two feature parameter sets are used in evaluation experiments: cepstral coefficients (Cep), and a combination of cepstral and regression coefficients (Cep and Δ Cep).

One-thousand-twenty-four codebook elements are produced by the LBG algorithm from the voices of four male speakers representing interspeaker variability. Each word is uttered by the four speakers, and each utterance is stored in the word dictionary after being converted into a sequence of codebook elements. Thus, four code sequences are stored per vocabulary word, one for each of the four reference speakers.

The test utterances consist of 100 city names uttered once by 20 male speakers who are different from the 4 reference speakers. Consequently, the number of test samples is $100 \times 20 = 2000$. Word utterances spoken

separately before the test utterances are used for the adaptation process for each speaker.

The effectiveness of the adaptation algorithm under various conditions is evaluated by measuring the spectral distortion (difference) between test utterances and reference templates (sequences of adapted codebook elements) and by recognition accuracies. The spectral distortion for each test utterance is defined as

$$D = \min_{k=1,4} \left(\sum_{i=1}^I d(i, j(i), k) / I \right), \quad (5)$$

where $d(i, j(i), k)$ indicates the distortion at the i th frame when the utterance is compared to the k th adapted reference template (code sequence) of the same word ($k = 1, \dots, 4$), I is the number of frames in the test utterance, and $j(i)$ indicates the frame number of the reference template which is matched to the i th frame of the test utterance by a dynamic time-warping algorithm.

The distortion $d(i, j(i), k)$ is calculated as the Euclidean distance between feature parameter vectors. When using both cepstral and regression coefficients, regression coefficients are appropriately weighted to balance the variance of the two sets of parameters [1]. The distortion at each frame, i , is averaged over I frames to produce overall distortion from the k th adapted reference template sequence. The minimum value, D , is then selected from the four overall distortions associated with the reference templates.

The distortion, D , and the recognition accuracy, respectively, averaged over all test utterances and test speakers, are then used for the evaluation.

B. Experimental Results

Spectral distortion after adaptation, which is normalized by the distortion before the adaptation, is shown in Fig. 3 as a function of the number of clusters. The parameter α , which controls continuity at the codebook shift, is set at 1 based on preliminary experiments. The figure shows the results for Methods I and II, and the parameter sets of (Cep) and (Cep and Δ Cep). For Method I, the adaptation is hierarchically performed by increasing the number of clusters, whereas for Method II, the adaptation is performed directly for the various cluster numbers. The number of training words is set at 10, and the results for two training word sets are averaged.

These results indicate that in Method I distortion decreases as adaptation progresses. On the other hand, with Method II, distortion has no simple relationship with the number of clusters and does not necessarily decrease with the number of clusters. The hierarchical algorithm effectively decreases distortion. When the number of clusters is larger than 16, better performances can be obtained by Method I than by Method II. The figure also indicates that the effect of adaptation for (Cep) is larger than for (Cep and Δ Cep).

Recognition error rates for (Cep) before and after adaptation are indicated in Fig. 4. The number of clusters is set at 8, 32, and 128. The results correspond well to the spectral distortion results presented in Fig. 3.

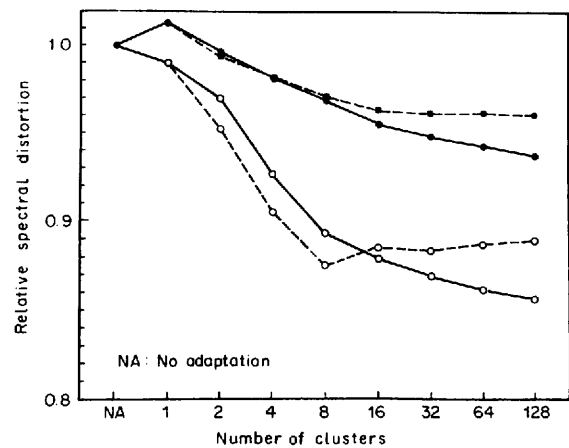


Fig. 3. Change in spectral distortion between input speech and word reference templates as speaker adaptation progresses. — Method I, --- Method II, ● Cep and △ Cep, ○ Cep.

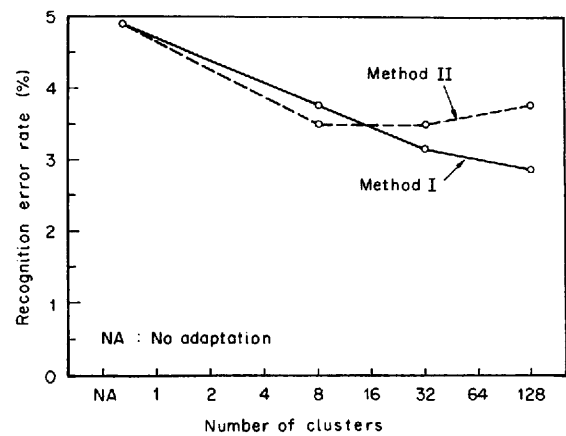


Fig. 4. Error rates for recognition using the parameter set (Cep) as a function of the number of clusters.

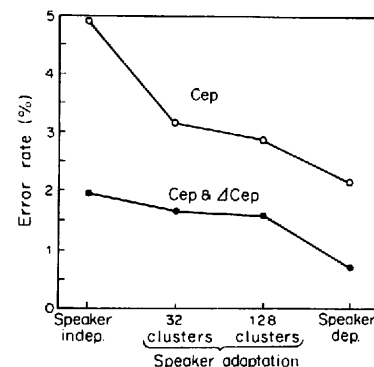


Fig. 5. Recognition error rates under four experimental conditions (Method I).

Recognition error rates under four experimental conditions are compared in Fig. 5. These conditions are 1) no adaptation (speaker-independent recognition), 2) adaptation up to 32 clusters by Method I, 3) adaptation up to 128 clusters by Method I, and 4) speaker-dependent recognition. In the last condition, all reference templates are constructed for each speaker using all vocabulary words uttered separately from the test utterances.

The experimental results show that, when (Cep) is used as a feature vector, the error rate is decreased from 4.9 to

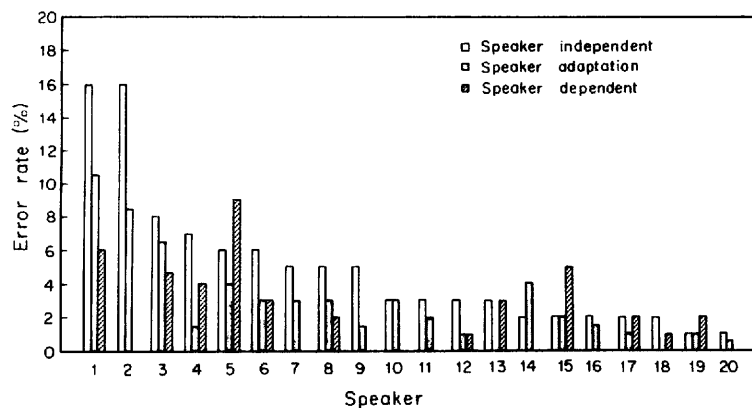


Fig. 6. Distribution of recognition error rates for 20 speakers for speaker-independent, speaker-adaptation, and speaker-dependent conditions [Method I, (Cep)]. Perfect recognition was achieved for eight speakers in speaker-dependent condition, and for two speakers in speaker-adaptation condition.

2.9 percent by adaptation up to 128 clusters. The error rate for the speaker-dependent case is 2.2 percent. This means that the adaptation method proposed in this paper is highly effective for reducing the error rate. When (Cep and Δ Cep) is used, the error rate is decreased from 2.0 to 1.6 percent by adaptation up to 128 clusters. Although the effectiveness using (Cep and Δ Cep) is less than using (Cep), it is still statistically significant.

Fig. 6 shows the distribution of recognition error rates for the 20 speakers with no adaptation, adaptation up to 128 clusters, and speaker-dependent recognition, with (Cep) and Method I used as a feature vector and an adaptation method, respectively. Comparing the speaker-adaptation results to the speaker-independent results, 16 speakers show a decrease in error rate, 3 show no improvement, and only 1 shows negative results. Therefore, it can be concluded that the adaptation method is fairly reliable for most speakers.

The figure also shows that the distribution of recognition error rates after adaptation does not correspond well with that for speaker-dependent recognition. One possible reason is that there is a large difference between the two training methods as follows. In speaker-dependent recognition, code sequences are adapted for each speaker using all vocabulary words uttered in training, while the universal codebook elements are maintained. On the other hand, with the speaker adaptation, the codebook elements are adapted, while the code sequences are maintained.

IV. DISCUSSION AND SUPPLEMENTARY EXPERIMENTS

A. Effects of the Number of Training Words

In order to investigate various important features of the adaptation method proposed in this paper, supplementary experiments were performed, in which Method I and (Cep) were used as an adaptation method and feature parameter set, respectively, unless otherwise stated.

In the previous experiments, the number of training words was fixed at 10. As the first supplementary experiment, the relationship between the number of training words and spectral distortion after adaptation was inves-

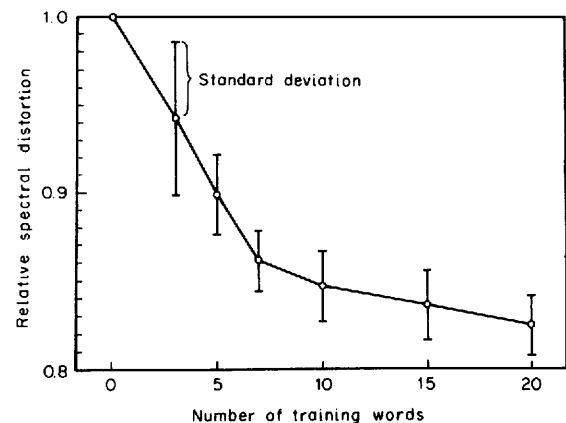


Fig. 7. Relative spectral distortion after speaker adaptation as a function of the number of training words.

tigated, with the maximum number of clusters being set at 128. For each condition of the number of training words, different training word sets with no common words between them were selected. Mean value and standard deviation for the intertraining set variation of spectral distortion were calculated for each condition and presented in Fig. 7. These results indicate that the mean distortion decreases rapidly when the number of training words is increased to 7, and after that it decreases only gradually.

B. Dependency of Adaptation Effect on Training Word Set

The results for standard deviation presented in Fig. 7 show that the variation of spectral distortion depending on training word sets is fairly constant irrespective of the number of training words when the number is larger than 5. When the number of training words is 10, the maximum and minimum distortions are 0.873 and 0.804. Based on these results, it can be concluded that the adaptation can be successfully achieved irrespective of the word content, if 7 or more distinctive words are used for training.

Table I shows the variation of normalized spectral distortion after adaptation for 10 training word sets, with the

TABLE I
VARIATION IN RELATIVE SPECTRAL DISTORTION AFTER ADAPTATION FOR
TEN TRAINING WORD SETS OF THREE WORDS

Training words	Relative distortion
(a) /saQporu/, /hakodate/, /asahikawa/	1.046
(b) /kufiro/, /aomori/, /hatfinohe/	0.945
(c) /morioka/, /seNdai/, /akita/	0.915
(d) /jamagata/, /fukufima/, /korijama/	0.895
(e) /iwaki/, /niigata/, /tojama/	0.958
(f) /kanazawa/, /fukui/, /nagano/	0.931
(g) /matsumoto/, /mito/, /hitatfi/	0.921
(h) /utsunomija/, /maebafu/, /lakasaki/	0.915
(i) /kawagoe/, /kawagutfi/, /urawa/	0.978
(j) /omija/, /tokorozawa/, /kofigaja/	0.919
Mean	0.942
Standard deviation	0.044

number of words being set at 3. There is no clear relationship between the adaptation effect and the words or phonemes included in the training sets.

C. Effects of Maintaining Continuity Between Adjacent Clusters

In the experiments so far, the parameter α for controlling the continuity between adjacent clusters was set at 1. The relationship between the adaptation effect and the α value was investigated by an experiment in which the number of training words was set at 10, and the maximum number of clusters was set at 128. It is clear from (3) that the smaller the value of α , the larger the equivalent size of the region in which continuity is maintained by smoothing the shifting amount for the codebook elements.

Results of the experiment are shown in Fig. 8. In this figure, $\alpha = \infty$ corresponds to the condition where the continuity is not considered and, therefore, the deviation vector for the centroid of the cluster in which the codebook element is included is used directly for shifting. The results indicate that maintaining continuity is crucial for adaptation: when continuity is not considered, the full adaptation effect is difficult to achieve. The optimum value of α is between $1/2$ and 1, or more precisely, around 0.7. Since variation of distortion as a function of α around the optimum position is small, α can be set at 1, for ease of calculation.

D. Dependency of Adaptation Effect on Feature Parameter Set

As described in Subsection III-B, the adaptation is less effective for (Cep and Δ Cep) than for (Cep) in both spectral distortion and recognition error rate. A supplementary experiment using (Cep and Δ Cep) was tried in which only cepstral coefficients were adapted. Regression coefficients

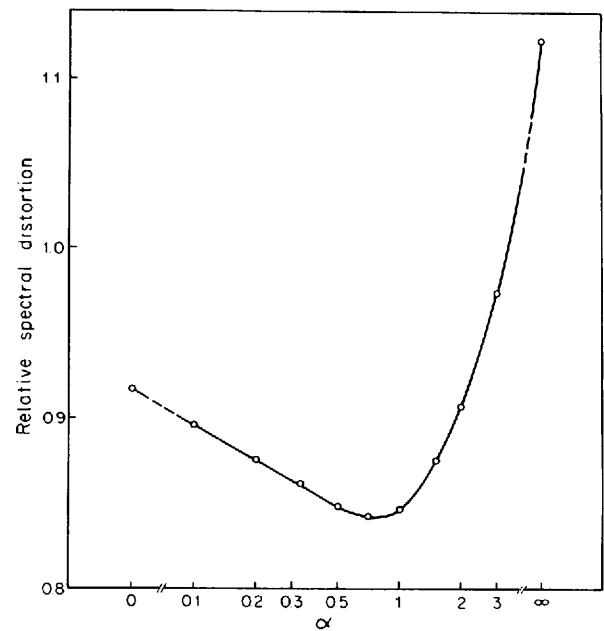


Fig. 8. Relative spectral distortion after speaker adaptation as a function of the smoothing coefficient, α .

in the universal codebook were maintained without adaptation. Results of the experiment indicate that the difference in recognition error rate between the previous and supplementary experiments is only 0.05 percent (one sample out of 2000). This means that the error rate remains essentially unchanged whether or not regression coefficients are adapted to each speaker.

Two factors may explain the small difference with and without adaptation for regression coefficients. First, the regression coefficients are relatively more speaker invariant than the instantaneous cepstral coefficients [15]. Second, the speaker-variation characteristics of regression coefficients are highly complicated and, therefore, not effectively adapted by the method proposed in this paper.

E. Combination of Methods I and II

When Method II is used and the number of training words is fixed at 10, the minimum spectral distortion can be obtained when the number of clusters is 8, as shown in Fig. 3. Distortion increases for both larger and smaller numbers of clusters. To examine the relationship between number of clusters and spectral distortion more precisely, additional experiments were performed with the number of training words set at 5 and 15.

The results of the experiments shown in Fig. 9 indicate that the number of clusters which produces the minimum distortion does not vary with the number of training words. Additionally, the variation of spectral distortion from changing the number of training words is larger for the smaller number of clusters. Fig. 3 further shows that Method II produces smaller distortion than Method I when the number of clusters is less than or equal to 8.

These results suggest that better adaptation performance might be achieved by setting the initial number of

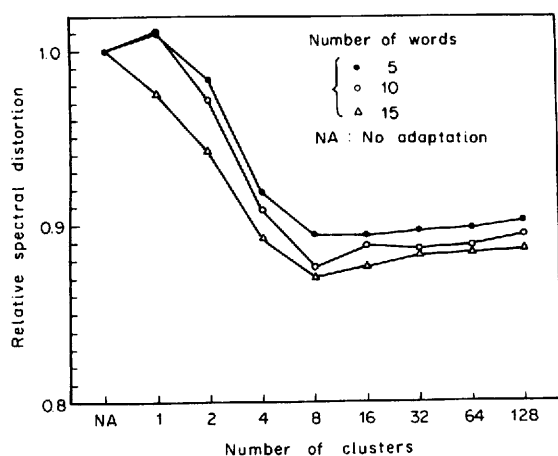


Fig. 9. Relationship between the number of clusters for adaptation using Method II and relative spectral distortion for three numbers of training words.

clusters larger than 1 for Method I when the number of training words is not large enough. When only a small number of training words are used, spectral matching between codebook elements and training utterances using the centroids obtained from global clustering having a very small number of clusters does not produce effective adaptation because the spectral distribution for the training utterance deviates greatly from the ideal spectral distribution for the speaker.

An additional experiment was then performed combining Methods I and II. Specifically, hierarchical adaptation was started from initial eight clusters. This method is referred to as Method III. Spectral distortion for this method as a function of the number of clusters is shown in Fig. 10 along with the aforementioned results for Methods I and II. This result indicates that spectral distortion is less for Method III than for Method I.

A recognition experiment was also performed using Method III, setting the maximum number of clusters at 128. It was confirmed that the mean recognition error rate for (Cep) can be slightly reduced by using Method III. The error rate for Method III was 2.75 percent, whereas that for Method I was 2.90 percent.

F. Relationship Between Spectral Distortion and Error Rate

In this paper, effectiveness of adaptation has been mainly evaluated using spectral distortion, and several conditions have been examined through recognition experiments. In these investigations, we have assumed that the amount of spectral distortion is proportional to the recognition error rate. In order to justify this assumption, the distortion versus error rate relationship for various experimental conditions investigated in this paper is plotted in Fig. 11. Results for all the conditions under which recognition experiments reported in the previous sections have been performed are plotted in this figure together with some additional experimental results. These results show that the spectral distortion and recognition error rate are highly correlated.

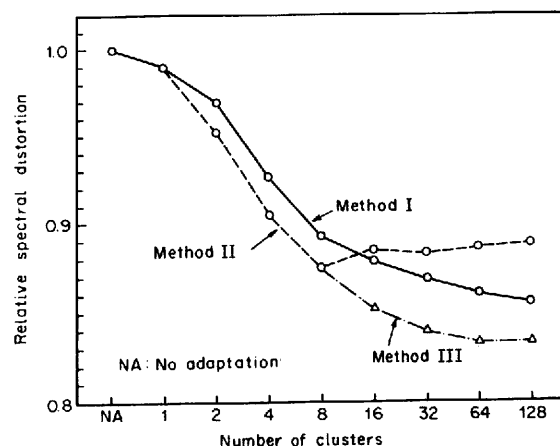


Fig. 10. Change in spectral distortion with the advance of speaker adaptation for Methods I, II, and III, when the feature vector consists of cepstral coefficients only (Cep).

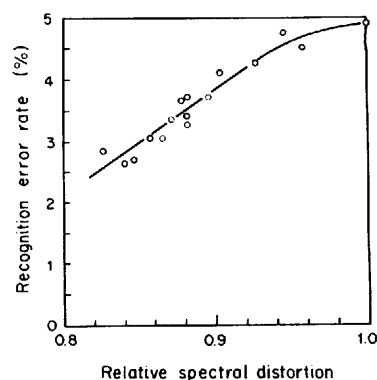


Fig. 11. Relationship between relative spectral distortion and recognition error rate for various conditions.

V. CONCLUSION

A new unsupervised speaker adaptation method for VQ-based speech recognition has been proposed based on the method for a segment vocoder. In this method, a set of spectra extracted from training utterance frames and the codebook entries are clustered hierarchically. Using the distances between centroids of the training frame clusters and the corresponding codebook clusters, adaptation is performed hierarchically from small to large numbers of clusters.

Recognition experiments were performed using utterances of 100 Japanese city names represented by sequences of a 10-dimensional LPC cepstral vector every 16 ms. An initial 1024 codebook entries were obtained from the utterances of four male speakers representing speaker variations. Ten arbitrary words recorded by each individual test speaker were used for adaptation. The results show that spectral distortion between test utterances and reference templates is decreased by the adaptation process, and that the mean word recognition error rate is reduced from 4.9 to 2.9 percent. Since the error rate for speaker-dependent recognition is 2.2 percent, the proposed adaptation method is highly effective.

The effectiveness of this method is also confirmed for the case when regression coefficients of cepstral coeffi-

cients and logarithmic energy sequences are combined with instantaneous cepstral coefficients as feature parameters, although the improvement (from 2.0 to 1.6 percent) is much smaller than the cepstrum-only case.

The effects of the number of training words, and of maintaining continuity between adjacent clusters, have also been investigated. The proposed method is general enough and is equally applicable to other speech recognition systems. It is also applicable to the spectral mapping approach for noisy speech recognition [16], [17], in which recognition systems adapt to speech modified by additive noise and by Lombard effect, based on a correspondence between clean and noisy signals.

ACKNOWLEDGMENT

The author wishes to thank F. K. Soong of AT&T Bell Laboratories for his comments and suggestions. The author wishes to acknowledge M. Honda, Y. Shiraki, and M. Sugiyama of NTT Basic Research Laboratories, and S. Sagayama of NTT Human Interface Laboratories for their valuable and stimulating discussions. The author also wishes to thank the reviewers for their valuable comments.

REFERENCES

- [1] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 52-59, 1986.
- [2] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336-349, 1979.
- [3] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1075-1105, 1983.
- [4] S. Furui, "A training procedure for isolated word recognition systems," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 129-136, 1980.
- [5] K. Shikano, K.-F. Li, and R. Reddy, "Speaker adaptation through vector quantization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986, Paper 49.5.
- [6] R. Schwartz, Y.-L. Chow, and F. Kubala, "Rapid speaker adaptation using a probabilistic spectral mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, 1987, Paper 15.3.
- [7] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, no. 2, pp. 175-184, 1952.
- [8] K. Kato and S. Furui, "Listener adaptability for individual voice in speech perception," *Trans. Committee Hearing Res., Acoust. Soc. Japan*, H-85-5, 1985.
- [9] S. Furui, "Learning and normalization of the talker differences in the recognition of spoken words," *Trans. Committee Speech Res., Acoust. Soc. Japan*, S-75-25, 1975.
- [10] —, "On the method of separating the personal and phonemic information in speech wave," *Trans. Committee Speech Res., Acoust. Soc. Japan*, S-74-17, 1974.
- [11] M. Sugiyama, "Unsupervised speaker adaptation methods for vowel templates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986, Paper 49.4.
- [12] Y. Shiraki and M. Honda, "Speaker adaptations algorithms for segment vocoder," *Trans. Committee of Speech Res., Acoust. Soc. Japan*, SP87-67, 1987.
- [13] N. Sugamura, K. Shikano, and S. Furui, "Isolated word recognition using phoneme-like templates," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Boston, 1983, Paper 16.3.
- [14] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84-95, 1980.
- [15] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 1986, Paper 17.5.
- [16] D. Roe, "Speech recognition with a noise-adapting codebook," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, 1987, Paper 27.5.
- [17] B.-H. Juang and L. R. Rabiner, "Signal restoration by spectral mapping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Dallas, 1987, Paper 6.6.



Sadaoki Furui (M'79-SM'88) 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 Communications Laboratories, Nippon Telegraph and Telephone Corporation, in 1970, he studied the analysis of speaker characterizing information in the speech wave, its application to speaker recognition as well as interspeaker normalization in speech recognition, the vector-quantization-based speech recognition algorithm, and the analysis of speech perception mechanism. He is currently the Executive Manager of Speech and Acoustics Laboratory at NTT Human Interface Laboratories. From December 1978 to December 1979 he was with the Staff of the Acoustics Research Department at Bell Laboratories, Murray Hill, NJ, as a visiting researcher working on speaker verification.

Dr. Furui is a member of the Acoustics, Speech, and Signal Processing Society Technical Committee on Speech, the Acoustical Society of Japan, and the Institute of Electronics, Information, and Communication Engineers of Japan.