IEEE 1989 International Conference on
ACOUSTICS, SPEECH,
AND SIGNAL PROCESSING
MAY 23-26 1989 GLASGOW, SCOTLAND, U.K.

UNSUPERVISED SPEAKER ADAPTATION METHOD BASED ON HIERARCHICAL SPECTRAL CLUSTERING

Sadaoki Furui
NTT Human Interface Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper proposes a new automatic speaker adaptation method for speech recognition, in which a small amount of training material of unspecified text can be used. This method is easily applicable to VQ-based speech recognition systems where each word is represented as multiple sequences of codebook entries. In the adaptation algorithm, either the codebook is modified for each new speaker or input speech spectra are adapted to the codebook, thereby using codebook sequences universally for all speakers. The important feature 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. Results of recognition experiments indicate that the proposed adaptation method is highly effective. Possible variations using this method are also presented.

I . INTRODUCTION

Recently speaker-independent speech recognition systems such as those using HMM techniques have been actively investigated, and recognition accuracy has been significantly improved [1] [2]. However, to obtain high performance for every speaker, in other words, to cope with the so-called "sheep and goats phenomenon", it is necessary to combine a speaker adaptation mechanism with these systems. 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. 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. Recognition errors produce inappropriate adaptation in these methods.

It is desirable that the system be used like a speaker-independent system, which requires no additional training utterance from each speaker, and that the system automatically adapts to the speaker's voice based on the current speech characteristics. It is also desirable that the system automatically adapts to a new speaker, without being explicitly informed of a change of speakers ("self-adapting system"). This paper investigates a new unsupervised speaker adaptation method that addresses these needs. Such systems are also expected to adapt to variations such as different telephone sets and microphones, different background noise conditions and room acoustics, and variable bandwidth limitations of different transmission systems.

II SPEAKER ADAPTATION ALGORITHMS FOR VQ-BASED MULTIPLE-TEMPLATE SPEECH RECOGNITION SYSTEMS

2. 1 Overview

In a variety of speaker adaptation algorithms for speech recognition systems investigated in this paper, speech is represented by time sequences of VQ (vector quantization) codebook elements created by clustering of a feature vector set. The codebook, which represents instantaneous and transitional spectra, 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 multiple templates method, which is a successful method for speaker-independent speech recognition, assures moderate recognition performance for new speakers at the beginning of the adaptation process.

The principle of the unsupervised speaker adaptation method proposed here is based on the method proposed by Shiraki et al. [3] for adapting the segment vocoder codebook to a new speaker. This paper investigates the feasibility of applying the speaker adaptation method for the segment vocoder to speech recognition and examines various important features of this method [4].

2. 2 Recognition System Structures Including Speaker Adaptation

The method proposed here can be applied to every kind of VQ-based speech recognition system, from isolated word to continuous speech recognition. This paper evaluates the effectiveness of this method, using a SPLIT isolated word recognition system [5]. Figure 1 shows block diagrams of the isolated word recognition systems including the three speaker adaptation algorithms investigated in this paper. A word dictionary (a set of multiple VQ-code sequences) is commonly used for all speakers, and either a codebook is adapted to each speaker or input speech spectra are adapted to the codebook. Four speakers ($N=4$) representing inter-speaker variability are selected, and the voices of these speakers are used for creating an initial universal codebook and word dictionary. The universal codebook is, therefore, a union of four speakers' individual codebooks. Four code-sequences are stored per vocabulary word, one for each of the four speakers. These procedures are performed for male and female speakers separately.

Structure (a) is the method in which the universal codebook is directly used, and speaker adaptation is performed based on the relationship between input speech spectra and universal codebook elements.

In structure (b), codebooks for each reference speaker, subsets of the universal codebook, are separately used for speaker adaptation. First, a codebook which produces minimum average VQ distortion for input speech is selected. Then, input speech spectra are adapted to the selected codebook. Since universal codebook elements are not necessarily used as the same phoneme with different speakers, adaptation of a selected codebook to the input speech is likely to produce undesirable modifications in multiple templates associated with other reference speakers. Therefore, only the method in which input speech is adapted to the selected codebook is employed in this structure.

Structure (b) is expected to be better than structure (a) when the recognition system must adapt to a wide range of both male and female speakers. When a single universal codebook is used for the wide range of speakers, the distribution and overlap in the spectra of various phonemes are thought to be too large for effective adaptation of individual variation in each phoneme. When structure (b) is used, the system is capable of adapting to a wide range of speakers, since input speech is adapted to a reference speaker who is the most similar to the input speaker.

Structure (c) is the method in which the codebooks for reference speakers that comprise the universal codebook are adapted or mapped to a codebook for one reference speaker before the adaptation. Since the distribution of universal codebook elements for each phoneme are reduced by this process, it is expected that the universal codebook can be directly used for the adaptation without selecting the most similar individual's codebook. The word dictionary is re-structured using the condensed universal codebook.

The most important point of these adaptation algorithms is in the use of hierarchical clustering of spectra in the training speech samples and in the VQ codebook. Based on hierarchical clustering, step-by-step adaptation from global to local individuality is performed using the deviation vectors between centroids of corresponding clusters. The adaptation can be performed using

CH2673-2/89/0000-0286 $1.00 © 1989 IEEE

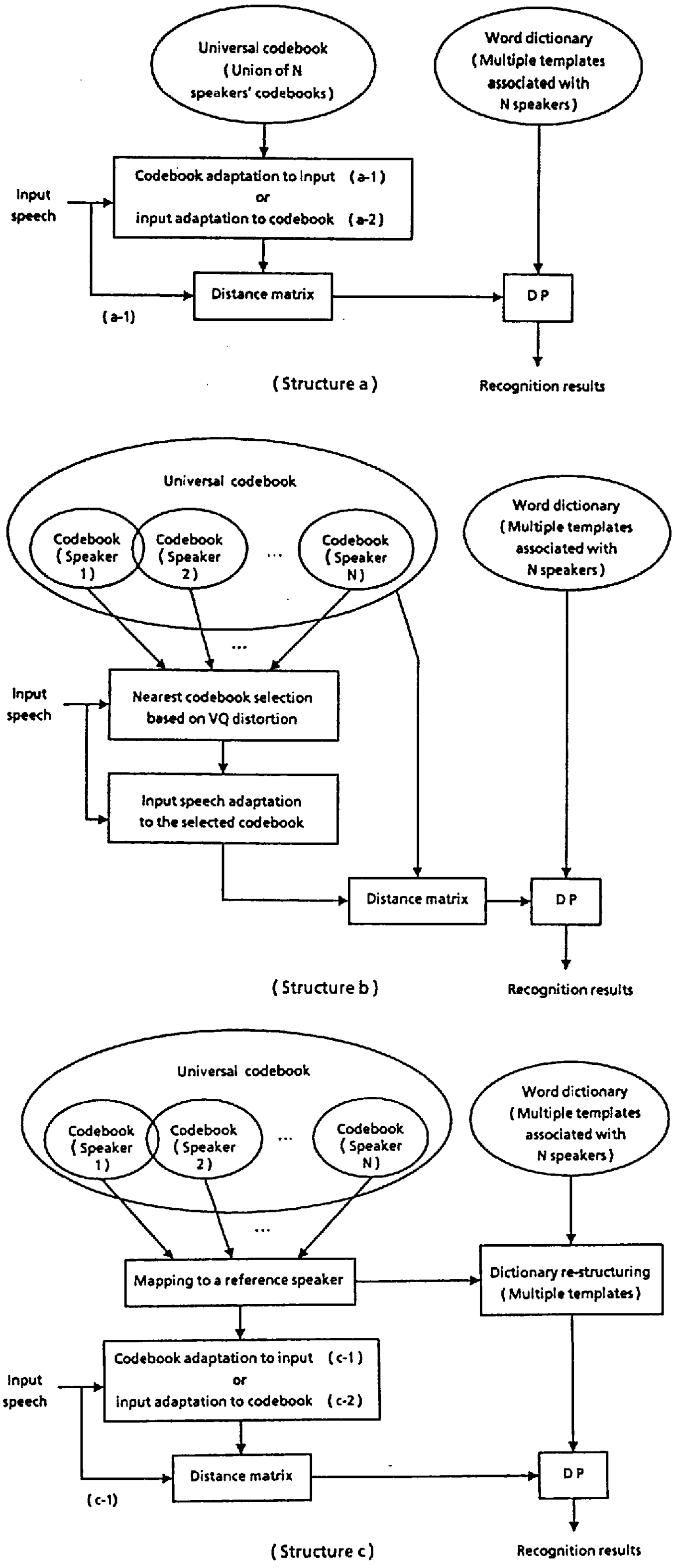

Fig. 1- Principal structures of word recognition systems with the speaker adaptation mechanism.

any word set or short sentence. The adaptation can even be done using unknown input utterances to be recognized. Experiments reported in this paper use word utterances for the sake of simplicity.

2. 3 Adapting the Codebook to Input Speech [3]

Procedures for adapting the codebook to input speech are as follows (see Fig. 2). The size of the codebook is maintained throughout the adaptation process.

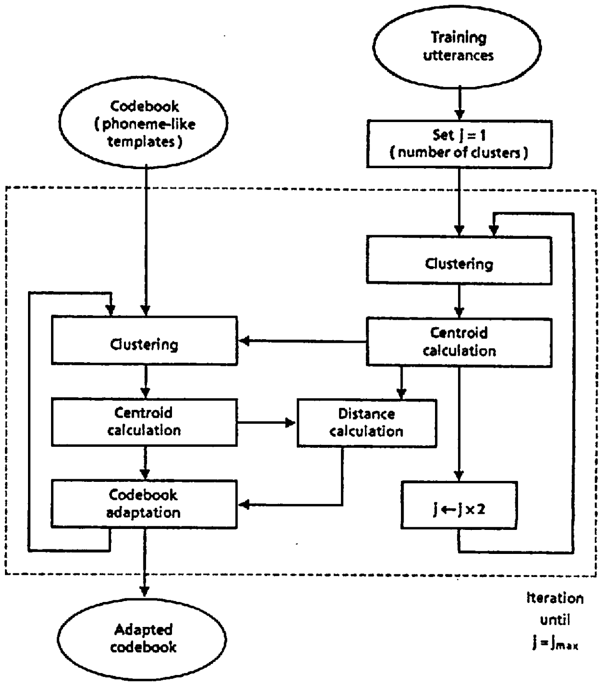

Fig. 2- Block diagram of codebook adaptation to the input (training) speech spectra.

(1) The initial number of clusters is set at 1, and the maximum is set at j_{max}.
(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.
(5) The centroid of each training utterance cluster is calculated, and codebook elements are clustered by assigning to the closest training utterance clusters.
(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 Subsec. 2.4)
(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) to (8). Input utterances are then recognized by the SPLIT method, using the adapted codebook elements.

This adaptation method using hierarchical clustering is hereafter referred to as Method Ⅰ. To evaluate the effectiveness of this method it is compared with 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) are performed under the condition of $j=j_{max}$, without iteration.

2. 4 Codebook Element Shift Maintaining Continuity Between Adjacent Clusters [3]

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 the corresponding training speech cluster by v_m, the deviation vector between these two centroids is

$$p_m = v_m - u_m \quad (m=1,\cdots,j) \tag{1}$$

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:

$$\Delta_i=(\sum_{m=1}^{j} w_{im}p_m)/(\sum_{m=1}^{j} w_{im}) \qquad (2)$$

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

$$w_{im}=1/\parallel c_i-u_m \parallel^{\alpha} \qquad (3)$$

The codebook element c_i is then shifted to c_i' as

$$c_i'=c_i+\Delta_i \qquad (4)$$

2.5 Procedures for Adapting Input Speech to the Codebook

The procedures for adapting input speech spectra to the codebook are similar to those for adapting the codebook to input speech spectra described in the previous subsections. Here, codebook elements are clustered first, and these centroids are used for clustering the training utterance spectra. The deviation vectors between codebook element cluster centroids and corresponding training utterance cluster centroids are produced. Using these deviation vectors, training utterance spectra are shifted, while continuity between adjacent clusters is maintained. The number of clusters is initially set at 1, and doubled at each repetition of these procedures. Training utterance spectra are thus hierarchically adapted to the codebook.

In this method, original training utterances are stored in the memory and used later for obtaining deviation vectors between corresponding training utterance spectra before and after the adaptation for each training utterance frame. For each frame of input speech, the closest training utterance frame is selected, and the input spectrum is shifted (adapted) using the corresponding deviation vector. In this adaptation procedure, it is unnecessary to cluster the codebook elements every time, since the codebook elements are maintained for all speakers. Therefore, the amount of calculation can be reduced by advanced storage of the centroids of codebook element clusters for every hierarchical clustering stage and by using these centroids in the adaptation process.

Ⅲ. RECOGNITION EXPERIMENTS

3.1 Recognition Task and Feature Extraction

One hundred Japanese city name utterances are used in isolated word recognition experiments to evaluate the adaptation algorithms described in the previous section. Speech signals are sampled at 8kHz and spectrally represented by the sequences of the first 10 LPC cepstral coefficients and the logarithmic energy at 16-ms frame intervals. The time derivatives for the cepstral coefficient and the logarithmic energy sequences are approximated by first order regression coefficients over a finite length window of 88 ms, centered around the current frame [6]. Two feature parameter sets are used in evaluation experiments: cepstral coefficients (Cep), and a combination of cepstral and regression coefficients (Cep & ΔCep). The size of the VQ codebook is set at 1024, and codebook elements are produced by the LBG algorithm.

The test utterances consist of 100 city names uttered once by 20 male speakers and 16 female speakers, all who are different from the reference speakers. Consequently, the numbers of test samples are 2000 and 1600, respectively. Word utterances spoken separately before the test utterances are used for the adaptation process for each speaker. The effectiveness of the adaptation algorithms under various conditions is evaluated by measuring the spectral distortion (difference) between test utterances and reference templates (sequences of codebook elements) and by the recognition accuracy. The distortion and the recognition accuracy, respectively averaged over all test utterances and test speakers, are used for the evaluation.

3.2 Experimental Results for Structure (a-1)

Structure (a-1), in which the universal codebook is adapted to input speech, is selected as a basic method, and various kinds of experiments are performed. Spectral distortion after the adaptation between male and female speakers, 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. 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.

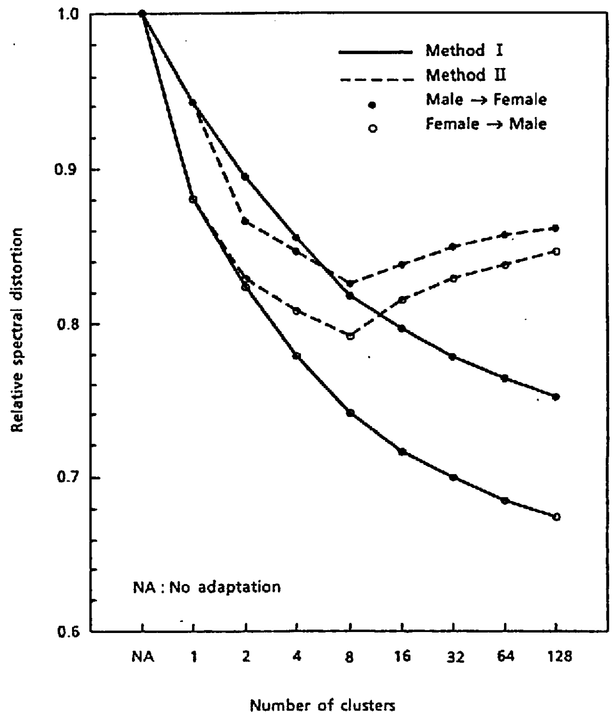

Fig. 3- Change in spectral distortion between input speech and word reference templates as speaker adaptation progresses (adaptation between male and female speakers).

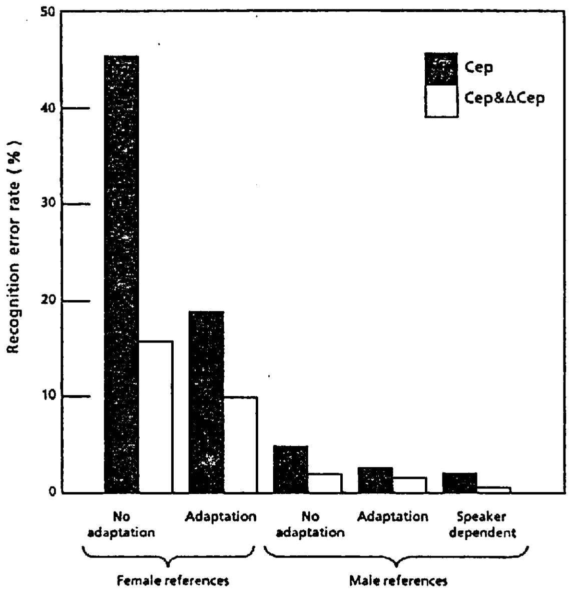

Fig. 4- Recognition error rates under various experimental conditions.

This result indicates 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. In the case of large spectral differences between reference and input speech such as those between male and female voices, there is a large difference between the minimum distortion values achievable by Method Ⅰ and Method Ⅱ. Based on this result, Method Ⅱ is removed from the investigation hereafter.

Recognition error rates for male speech using male or female references under three experimental conditions are compared in Fig. 4. These conditions are (1) no adaptation (speaker-independent recognition), (2) speaker adaptation up to 128 clusters, and (3) 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 the adaptation method proposed in this paper is highly effective for reducing the error rate. When (Cep) is used as a feature vector and male voices are used as references, the mean error rate is decreased from 4.9% to 2.9% by the speaker adaptation procedure. The error rate for the speaker-dependent case is 2.2%. When (Cep & ΔCep) is used, the error rate is decreased from 2.0% to 1.6% by the adaptation. Although the effectiveness using (Cep & ΔCep) is less than using (Cep), it is still statistically significant.

The relationship between the number of training words and spectral distortion after adaptation was also investigated. The experimental results presented in Fig. 5 indicate that the distortion decreases rapidly when the number of training words is increased to 7, and after that, decreases only gradually. It is also indicated that the variation of spectral distortion depending on training word sets is fairly constant regardless of the number of training words, when the number is larger than 5. It can be concluded that the adaptation can be achieved regardless of word content, if 7 or more distinctive words are used for training.

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. Results of the experiments for adaptation 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 both for male and female speakers, and variation of distortion as a function of α around the optimum position is small.

3. 3 Recognition Experiments by Adapting Input Speech to the Codebook

Recognition experiments are performed using the structures of (a-2) and (b) in which input speech spectra are adapted to the codebook. Adaptation is performed within male voices and (Cep) is used as the feature parameter set. In case of the structure (b), two kinds of experiments are tried: the codebook to which input speech spectra are adapted is selected arbitrarily from the universal codebook, or the codebook closest to the input speech is selected based on VQ distortion. Table 1 shows the experimental results including the error rate before the adaptation.

These results indicate that adaptation to a single speaker codebook produces better performance than adaptation to a universal codebook consisting of multiple speaker codebooks. Adaptation to the codebook closest to the input speech produces better results than using an arbitrarily selected individual codebook. The recognition error rate for the last method is, however, almost equal to that obtained using the structure (a-1), in which the universal codebook is adapted to the input speech. Structure (b) is expected to be advantageous when the system must deal with a much wider range of spectral variations including male and female voices. This remains to be investigated in the future.

3. 4 Recognition Experiments Using Structure (c)

Structure (c-1), in which condensed universal codebook is adapted to input speech, is also experimentally investigated using male voices. The recognition error rate obtained with this method is indicated in Table 2, comparing with the results before the daptation and the results using structure (a-1). These results indicate that the error rate can be reduced by condensing the universal codebook. The combination of the structures (a) and (c) is expected to produce better recognition performance than any other structure under more difficult circumstances having a wide range of speaker variations. This method also remains to be tried.

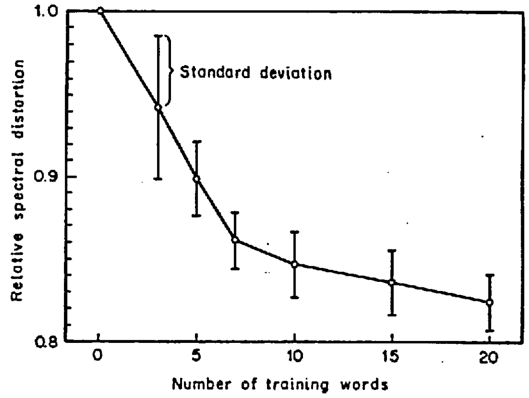

Fig. 5- Relative spectral distortion calculated using (Cep) after adaptation within male speakers as a function of the number of training words.

Table 1- Error rates for recognition experiments using methods of adapting input speech spectra to the codebook.

No adaptation	Structure a-2	Structure b (Random speaker selection)	Structure b (Nearest speaker selection)
4.90%	3.40%	3.15%	2.85%

Table 2- Error rates for recognition experiments with (structure c-1) or without (structure a-1) condensing the universal codebook.

No adaptation	Structure a-1	Structure c-1
4.90%	2.88%	2.60%

Ⅳ. CONCLUSION

New unsupervised speaker adaptation algorithms for VQ-based speech recognition have been proposed based on the method for a segment vocoder. In these algorithms, 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 is performed either by using a codebook adapted to the training utterance or by adapting every frame of input speech to the codebook, based on adaptation vectors obtained at the training stage. The experimental results show that recognition error rate is reduced significantly by the adaptation using 10 arbitrary training words.

Among the various variations, the procedure of adapting input speech to the most similar individual codebook selected from the multi-speaker universal codebook based on VQ distortion (structure b) and the procedure in which individual variations in the universal codebook are normalized (strucure c) are the most promising methods, because of their ability to adapt to a wide range of speaker variations. The proposed algorithms are general enough and are equally applicable to other speech recognition systems. They are also applicable to the spectral mapping approach for noisy speech recognition, in which recognition systems adapt to speech modified by additive noise and by the Lombard effect, based on the correspondence between clean and noisy signals.

ACKNOWLEDGMENT

The author wishes to acknowledge Masaaki Honda, Yoshinao Shiraki and Masahide Sugiyama of NTT Basic Research Laboratories and Shigeki Sagayama of NTT Human Interface Laboratories for their valuable and stimulating discussions.

REFERENCES

[1] 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., 62, 4, pp.1075-1105 (1983)

[2] K.-F.Lee and H.-W.Hon: "Large-Vocabulary speaker-independent continuous speech recognition using HMM", Proc.IEEE Int.Conf.Acoust., Speech, Signal Processing, New York, S3.7 (1988)

[3] Y.Shiraki and M.Honda: "Speaker adaptation algorithms for segment vocoder", Trans.Committee of Speech Res., Acoust.Soc.Jap., SP87-67 (1987)

[4] S.Furui: "Text-indepent speaker adaptation based on hierarchical spectral clustering", IEEE Workshop on Speech Recognition (1988)

[5] N.Sugamura, K.Shikano and S.Furui: "Isolated word recognition using phoneme-like templates", Proc.IEEE Int.Conf.Acoust., Speech, Signal Processing, Boston, 16.3 (1983)

[6] S.Furui: "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans., Acoust., Speech, Signal Processing, ASSP-34, 1, pp.52-59 (1986)