

論文 / 著書情報
Article / Book Information

Title	Speech Recognition Techniques Based on Context-Dependent/Independent Spectral Clustering
Author	SADAOKI FURUI
Journal/Book name	Proc. 2nd Symposium on Advanced Man-Machine Interface through Spoken Language, Vol. , No. , pp. 26-1-11
Issue date	1988, 11

SPEECH RECOGNITION TECHNIQUES BASED ON CONTEXT-DEPENDENT/INDEPENDENT SPECTRAL CLUSTERING

Sadaoki FURUI

NTT Human Interface Laboratories
Musashino-shi, Tokyo, Japan

ABSTRACT

This paper introduces recent speech recognition techniques, especially those for coping with the coarticulation and speaker variability problems, based on the concept of spectral space clustering. Spectral space clustering, that is, the vector quantization technique, was originally proposed to reduce memory capacity, transmission rate, and amount of computation. The concept of spectral space clustering has additionally created interesting and important techniques such as quantitative representation of coarticulation mechanisms based on phoneme environment clustering and automatic mapping of codebook entries for a new speaker. This concept is expected to continue to play an important role in speech recognition.

1. INTRODUCTION

How to deal with spectral variations due to coarticulation and individuality continues to be an important problem in speech recognition. Coarticulation is the phenomenon in which phoneme spectra vary with phoneme context, including adjacent and more distant phonemes. A large number of researchers are involved in HMM and neural net techniques. These techniques can cope with this problem to some extent, although they cannot completely solve it.

Clustering in the spectral space, that is, vector quantization (VQ) discretely represents the spectral space. This technique was proposed a long time ago to reduce memory capacity, transmission rate and the amount of computation, and it has since played important roles in speech coding and recognition areas. These kinds of quantitative reduction are achieved by this technique for two reasons. First, the spectral distribution of speech waves is continuous but not homogeneous. Second, the human auditory system is not sensitive to small spectral variations.

In addition to the quantitative reduction, the concept of spectral space clustering has produced various important techniques for solving the coarticulation and individuality problems. This paper introduces recent speech recognition techniques based on the concept of spectral space clustering, and describes future perspectives.

2. SPECTRAL SPACE CLUSTERING

Spectral space clustering is closely related to the problem of creating fundamental units for speech processing. It is desirable for the fundamental units to be characterized as not only linguistic units but also physical units (clusters), perceptual units and articulatory units for speech production, as shown in Fig. 1. Establishment of fundamental units is a basic and crucial problems in speech recognition, especially for realizing large vocabulary, continuous speech recognition systems.

Table 1 shows the categorization of spectral space clustering methods. These methods can be classified into two categories: whether linguistic (phonetic) labels for each spectral pattern are considered in the clustering procedure or not. The clustering methods can also be classified as static or dynamic. The static method simply clusters the observed instantaneous spectra or spectral time series. The dynamic method extracts time functions of categorized spectra which are supposed to be perceived by the human auditory system, by means of a spectral target-estimating mechanism based on spectral dynamics. It has been experimentally confirmed that the human auditory system is highly sensitive to spectral transition [1].

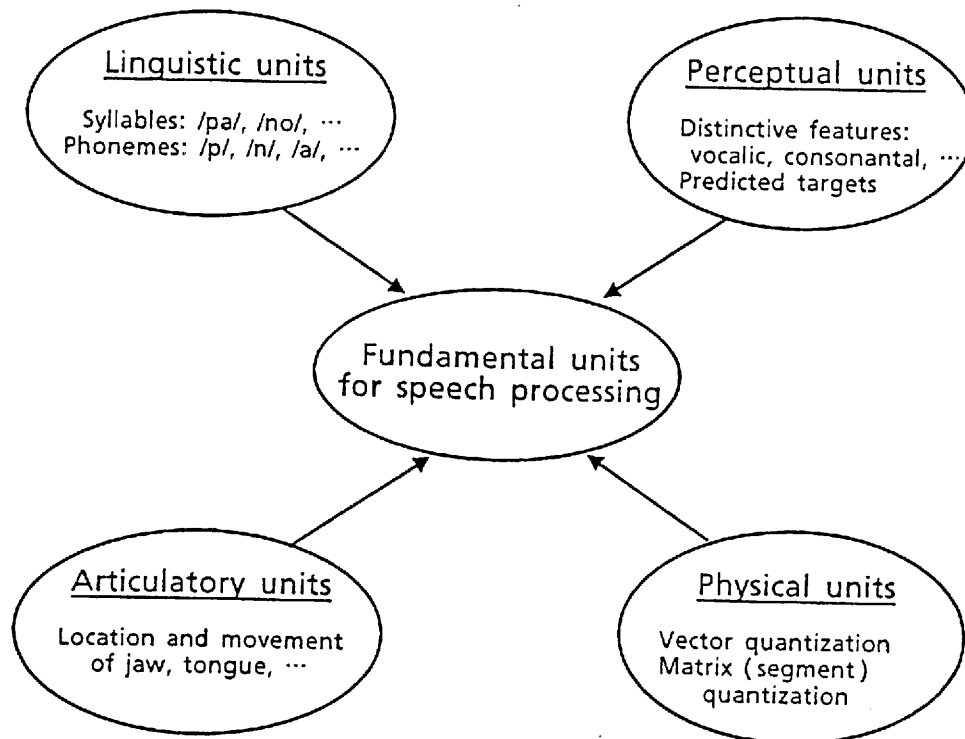


Fig.1 - Various aspects of fundamental units for speech processing.

Table 1 - Classification of spectral space clustering methods.

	Static clustering	Dynamic clustering
Labeled clustering (Top-down)	Phoneme segment model Phoneme environment clustering	
Un-labeled clustering (Bottom-up)	Vector quantization Matrix (segment) quantization	Target prediction

3. SPEECH RECOGNITION BY SPECTRAL SPACE CLUSTERING

3.1 Recognition Based on Unlabeled Clustering

Speech recognition methods using pseudo-phoneme reference template created by unlabeled clustering (VQ) can be traced back to research at IBM [2] and CMU [3] and to the SPLIT method by NTT [4][5]. In these systems, VQ was important in effectively reducing memory capacity and computation amount while maintaining high recognition accuracy. The methods of preselecting candidate words based on VQ distortion values calculated using word-specific codebooks were also proposed [6][7]. The VQ is also playing an important role in simplifying HMM-based methods. Recently a neural-net-based VQ technique has also been proposed [8].

Matrix quantization (MQ) or segment quantization (SQ) is the process of collectively encoding a time series of spectral patterns for a certain period using a single code. This technique has also been applied to speech recognition [9].

Since, in these methods, the codebooks are created based simply on physical features without considering linguistic labels, it is usually difficult to find a correspondence between each codebook element and phoneme class after clustering.

3.2 Recognition Based on Labeled Clustering

Recognition methods using spectral segments created through labeled spectral clustering have also been investigated [10] [11]. Each codebook element is therefore phonetically labeled in these methods. Figure 2 shows a block diagram of a word spotting system in which 750 labeled phoneme segments are used [10]. The variety of phoneme (allophone) segments is empirically determined by considering various kinds of coarticulation knowledge and spectral distance relationships between allophonic variations. Since each segment is labeled by phoneme environment comprising preceding and succeeding phonemes, each word reference template can be automatically created by concatenating phoneme segments according to orthographical phoneme representation. Then each word template is further expanded into a tree structure in accordance with phonetic rules to cope with phonetic variations such as long vowels and devocalized vowels. Word spotting is performed using continuous DP matching between continuous input speech and each tree-structured word template.

Phoneme segment generation using continuous speech as training samples, that is, spectral sequence clustering for continuous speech under contextual restriction, is accomplished by an automatic iterative labeling procedure utilizing the word spotting system. Initial segments for the iteration are given by manual labeling. Experimental results show that word spotting accuracy obtained by this system is almost equal to that obtained using word-based reference templates.

Continuous speech recognition using phoneme segments generated by applying heuristic rules to a labeled speech database and modeled by HMM, has also been investigated at several research laboratories [12][13]. Experimental results indicate that a context-dependent phoneme model that takes coarticulation (adjacent phonemes) into account is highly effective.

3.3 Phoneme Environment clustering

The factors affecting the variation of phoneme spectra, which we call phoneme environmental factors, include not only coarticulation but also many other phenomena, as shown in Fig. 3 [14]. The relationship between environmental factors and spectral variation has been investigated based on a mathematical clustering algorithm using a speech database, without relying on empirical knowledge. The database consists of spectral segments labeled by principal environmental factors. Clustering of phoneme spectral variation taking the environmental factors into account, in other words, the clustering of environmental factors, produces mapping between phoneme environment space and physical parameter space. The former abstract space is constructed by linguistic symbols, that is, phoneme symbols and environmental factors, and the latter physical space is constructed by spectral parameters, as shown in Fig. 4.

In the phoneme environment clustering algorithm, the phoneme environment space is divided into a specified number of subspaces (or clusters) so that the total distortion

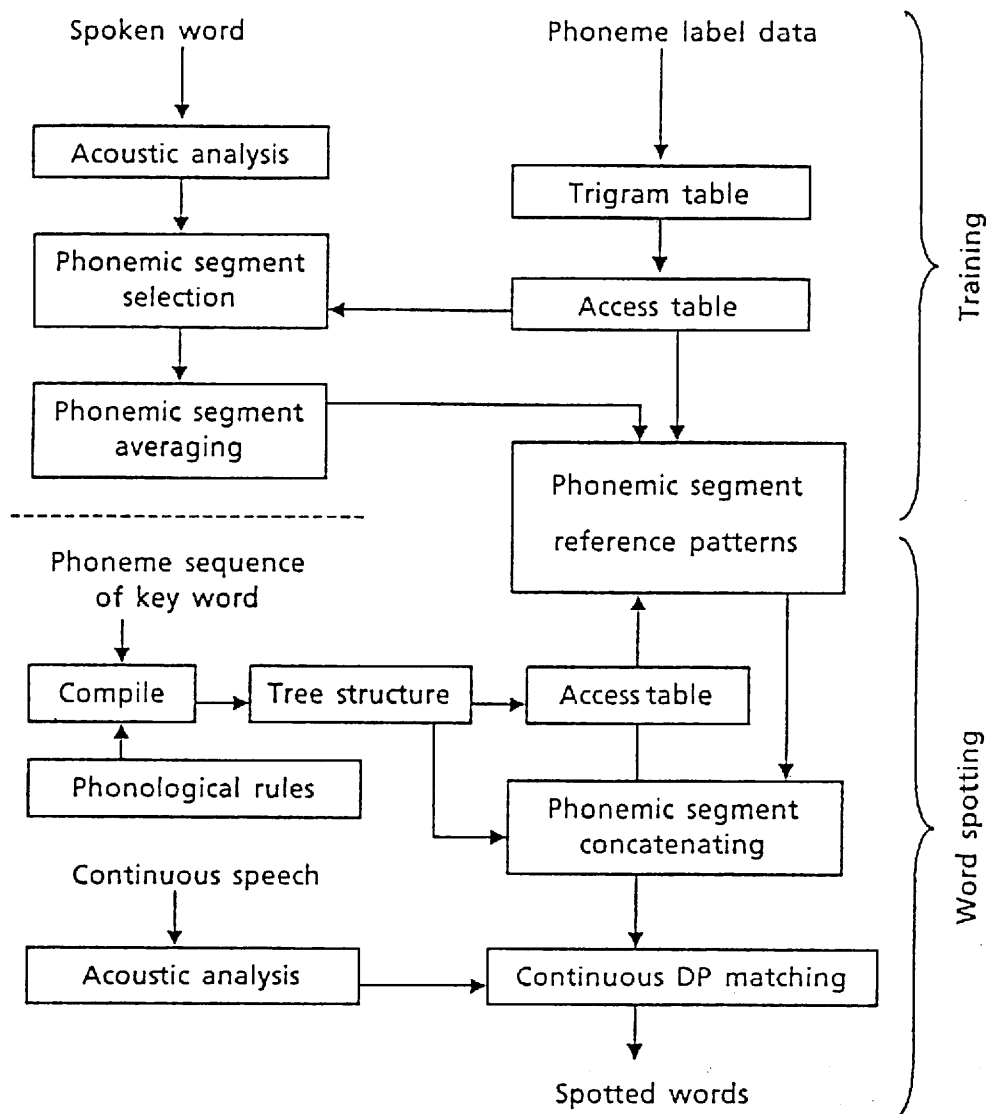


Fig.2 - Block diagram of a word spotting system using spectral segments created by labeled spectral clustering.

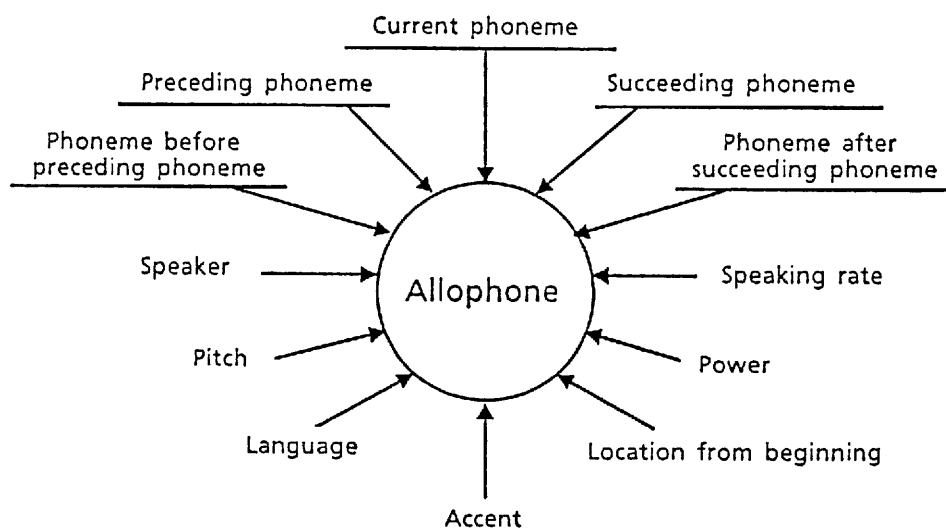


Fig.3 - Principal phoneme environmental factors.

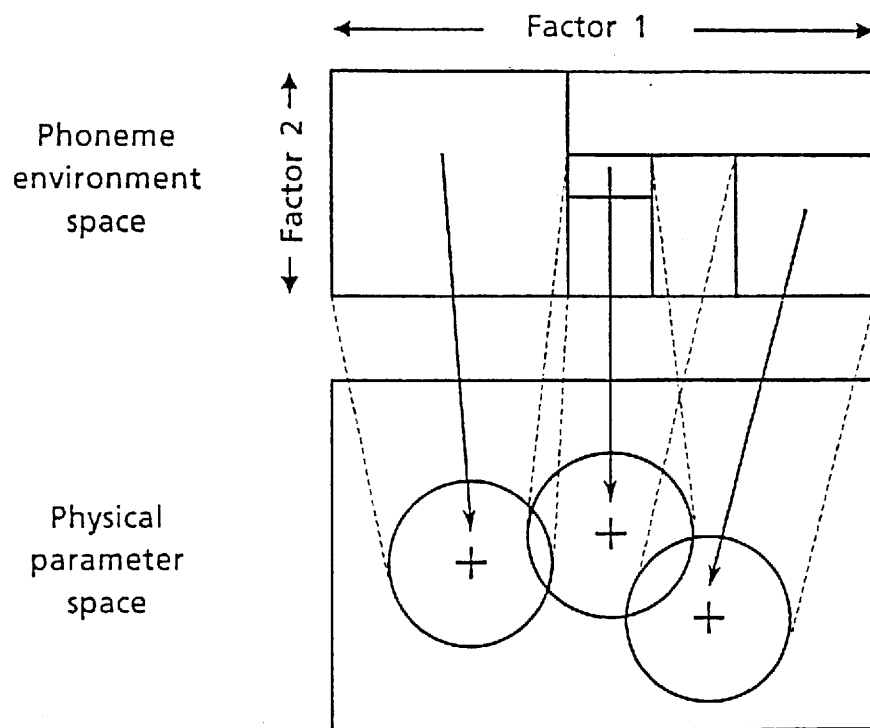


Fig.4 - Conceptual illustration of phoneme environment clustering.

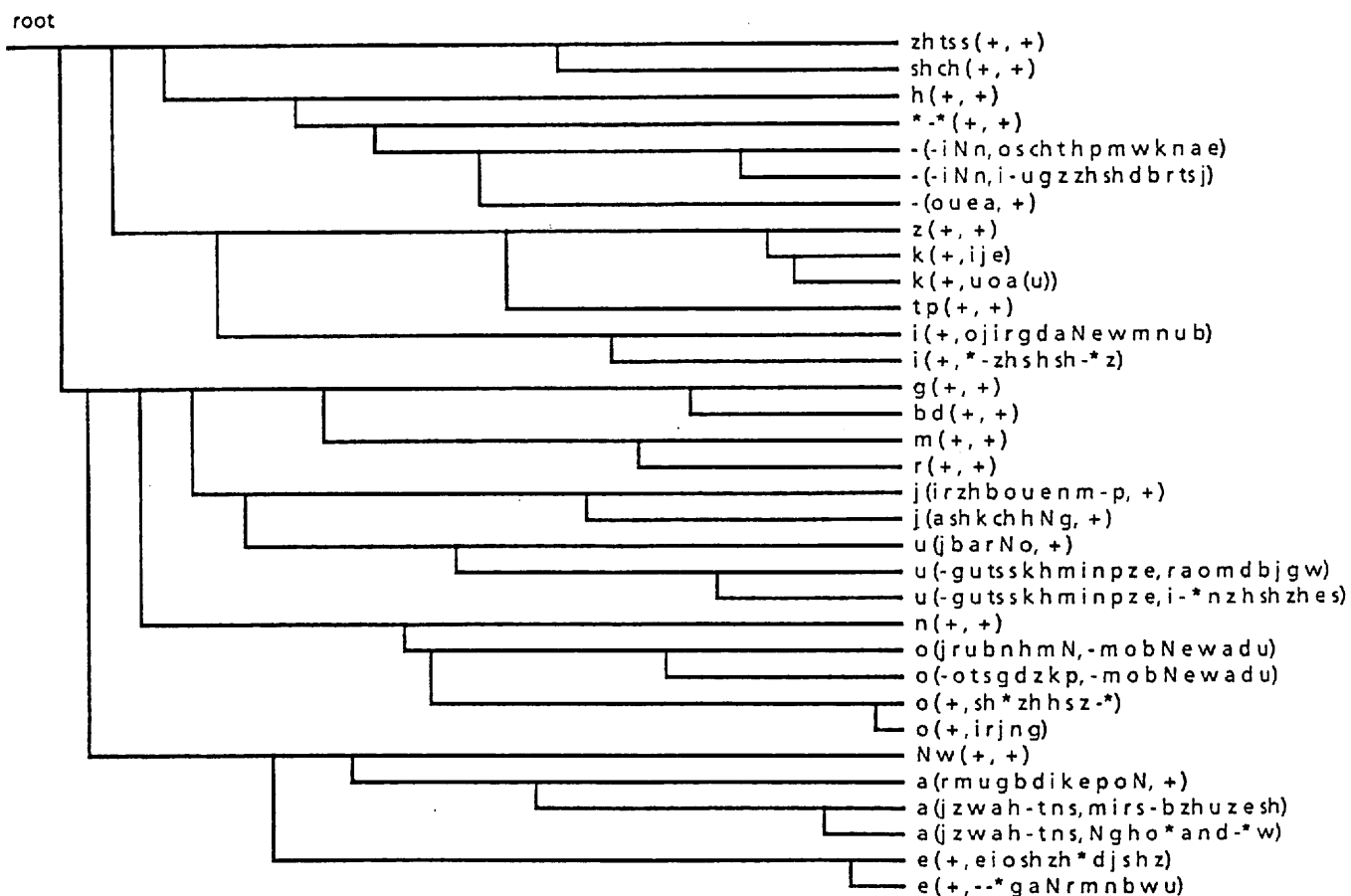


Fig.5 - Example of a phoneme environment subspace tree. Each leaf shows: center phoneme (preceding phonemes, succeeding phonemes).

resulting from mapping between subspaces and parameter space is minimized. The distortion associated with each subspace is defined as the distribution of mapped speech samples (images) in the parameter space. The actual procedure is realized by successive splitting of the phoneme environment subspace whose images have maximum distortion in the parameter space. If enough training database is available, the phoneme spectrum of any phoneme environment is expected to be estimated, within a certain tolerance, by this algorithm. Results of an experiment in which this algorithm is applied to a phoneme-based word recognition system have confirmed its effectiveness. Figure 5 shows an example of a phoneme environment subspace tree obtained by applying this method to a 216-word database.

4. SPEAKER ADAPTATION

Algorithms for adapting speech recognition systems to a new speaker have recently been proposed based on the concept of mapping VQ codebook elements. These algorithms can be classified as supervised, in which pre-determined training utterances are used to determine the mapping rules [15], and unsupervised, in which arbitrary utterances can be used for training [17][18]. In the supervised algorithm, codebook elements are adjusted to a new speaker based on rules obtained through DTW (dynamic time warping) or HMM fitting processes between the new speaker's utterances and reference templates.

In the unsupervised algorithm, which was originally proposed for a segment vocoder [19], a set of training utterance spectra and the codebook entries are clustered hierarchically. Based on the deviation vectors between centroids of the training clusters and the corresponding codebook clusters, adaptation is performed hierarchically from a small to a large number of clusters, that is, from phoneme-independent individuality to phoneme-dependent individuality. Spectral resolution of the adaptation process is improved accordingly. Figure 6 shows a block diagram of this method. The continuity of spectral space is taken into account in determining the shifting vector for each codebook entry, as shown in Fig. 7.

This method was evaluated by recognition experiments using utterances of 100 Japanese city names. Speech waves were analyzed using cepstrum and Δ cepstrum. Codebook and multiple word templates were constructed using utterances by four reference speakers. Figures 8 and 9 indicate a decrease in spectral distortion for a new speaker with increase in numbers of clusters using 10 arbitrary training words. The former figure shows results of adaptation between male speakers, and the latter shows results of adaptation between male and female speakers. Method I corresponds to the method described above, whereas Method II is a method in which clustering is performed directly on an specified number of clusters without hierarchical processing. Figure 8 shows results when using cepstrum only, and when using cepstrum and Δ cepstrum. Results of recognition experiments are shown in Fig. 10. These results indicate that the adaptation method is highly effective in reducing errors for a new speaker. The effectiveness of this method for Δ cepstrum is smaller than for cepstrum.

5. DYNAMIC CLUSTERING

The target of spectral transition is predicted and perceived based on spectral dynamic features in the human speech perception process. The target prediction process can be regarded as perceptual clustering of a spectral sequence. A mathematical model of this process was proposed, and its output was confirmed to correspond well with the results of perceptual experiments [20]. Figure 11 shows the phoneme sequences of input and output of this model for the utterance /kiai/. It clearly indicates that transitional periods are largely reduced, achieving clustering to the categorical targets.

6. DISCUSSION

Speech processing technology based on clustering in spectral space has been playing important roles in various aspects of speech recognition. This technology is expected to continue to create new techniques. Future problems include how to extract and represent the spectral dynamic features. Dynamic features are considered to be playing more

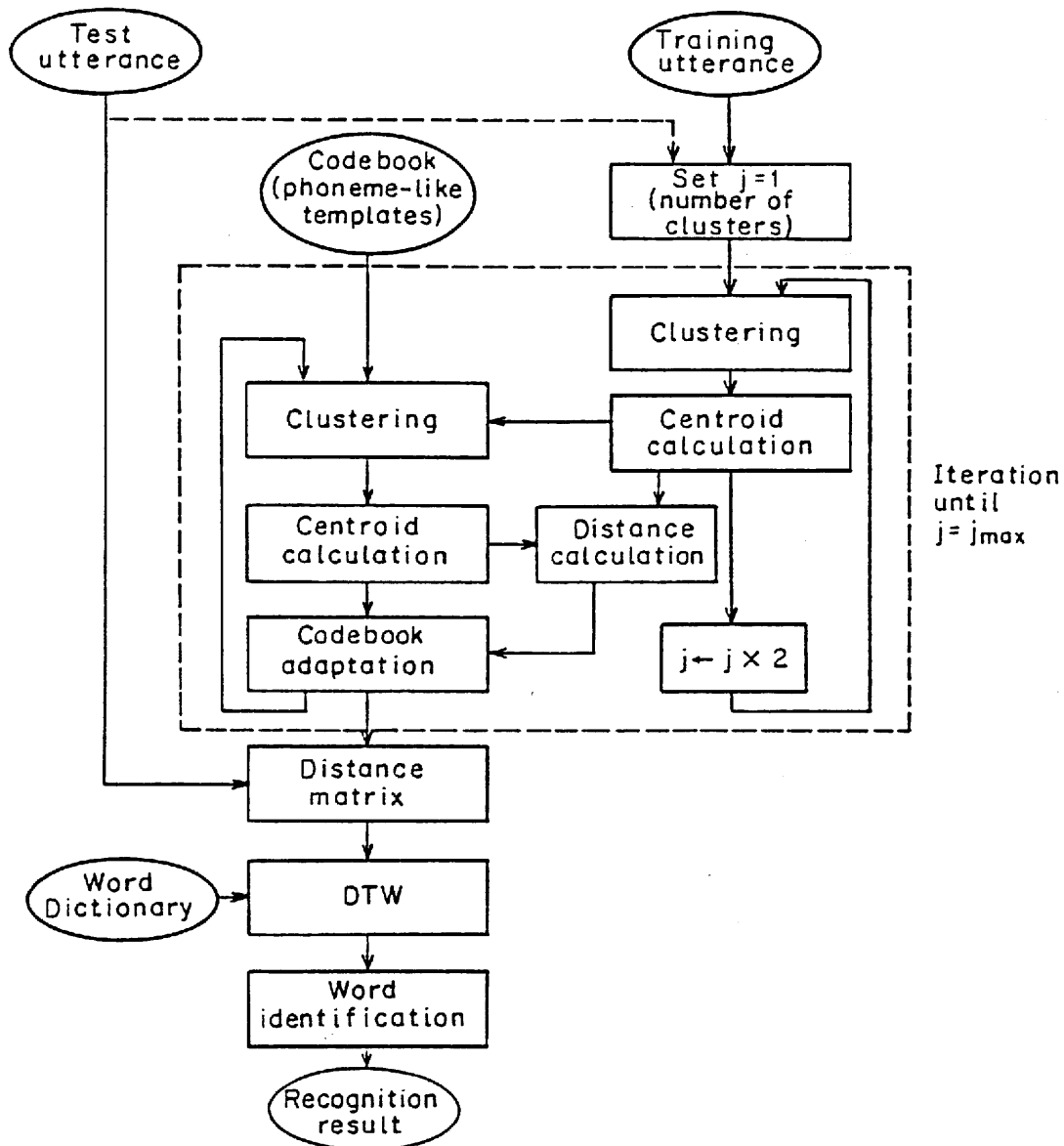


Fig.6 - Principal structure of a word recognition system containing unsupervised speaker adaptation mechanism.

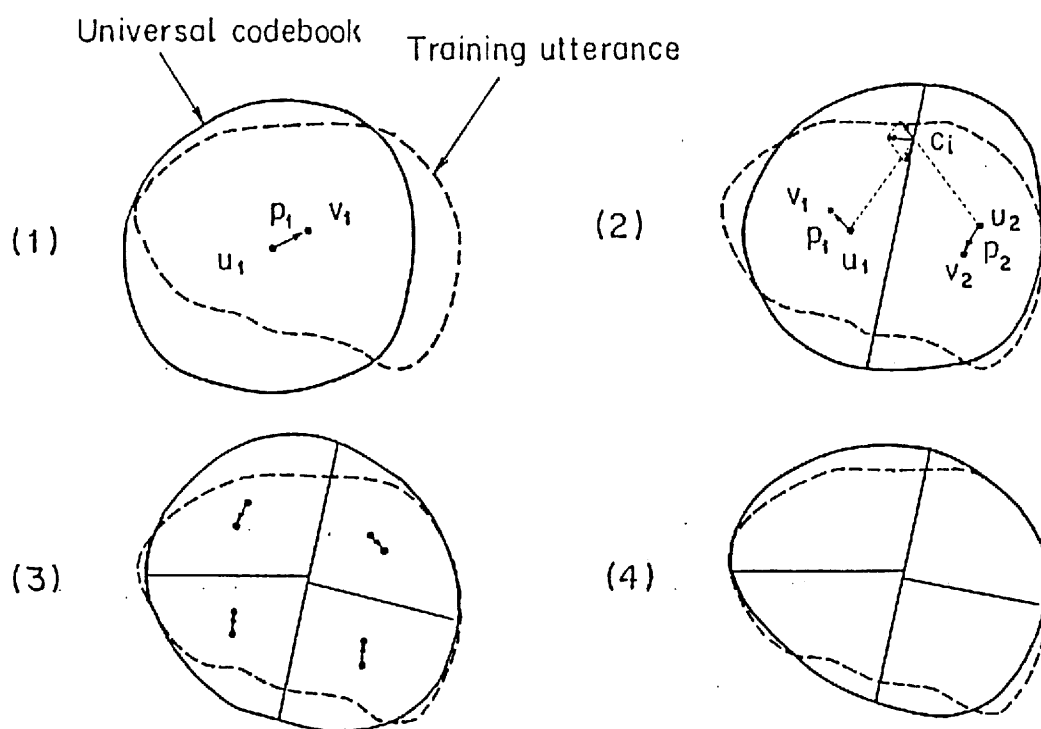


Fig.7 - Hierarchical codebook adaptation algorithm for maintaining continuity in spectral space.

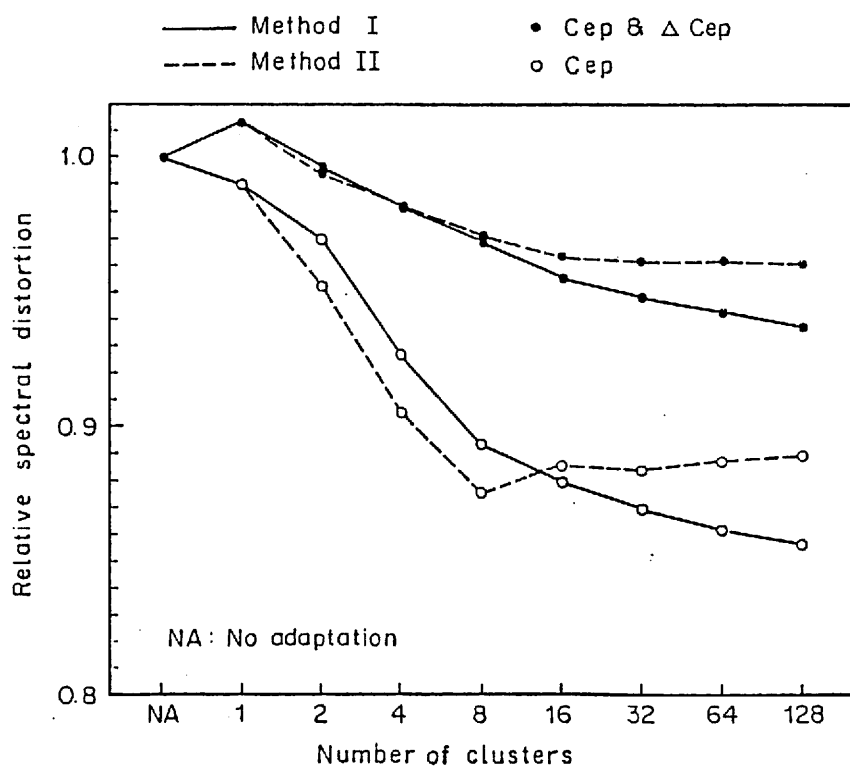


Fig.8 - Change in spectral distortion between input speech and word reference templates with speaker adaptation (adaptation between male speakers).

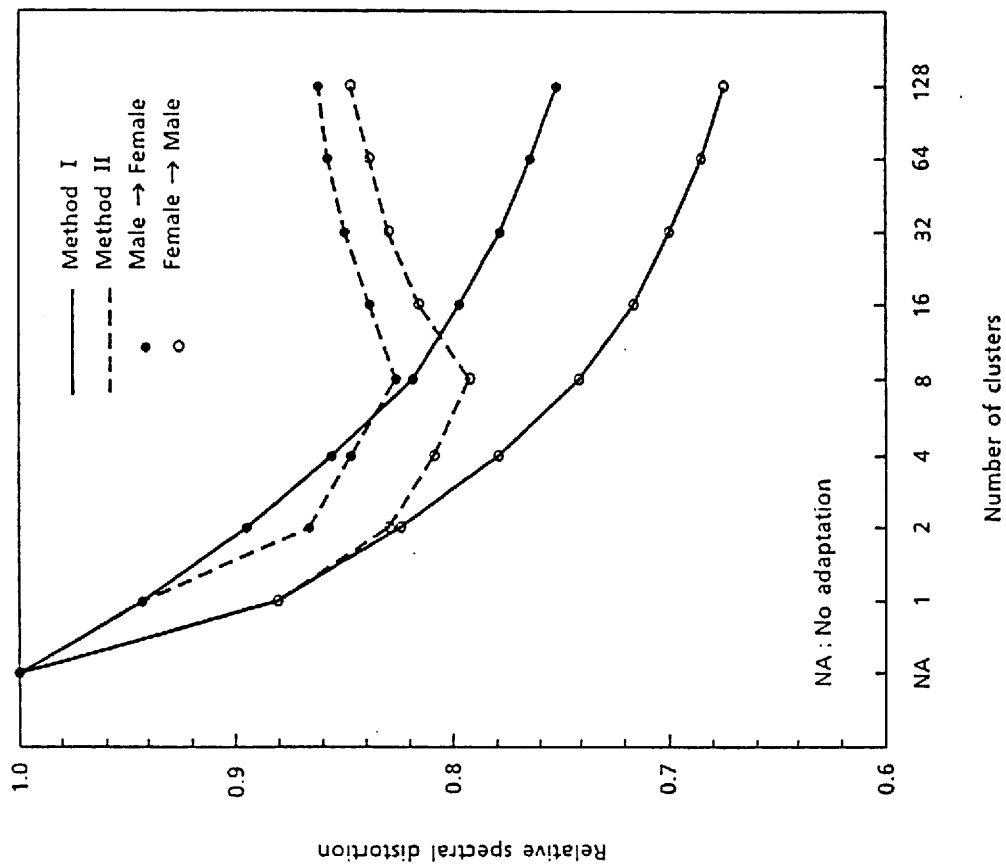


Fig.9 - Change in spectral distortion between input speech and word reference templates with speaker adaptation (adaptation between male and female speakers).

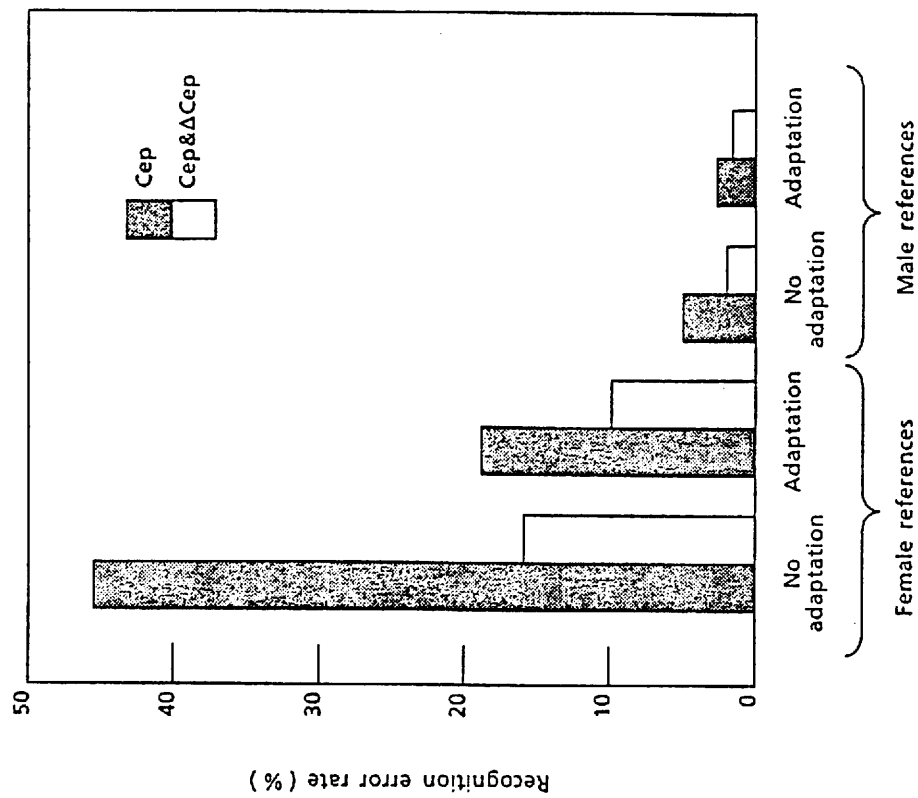


Fig.10 - Recognition error rates for male voice under four experimental conditions.

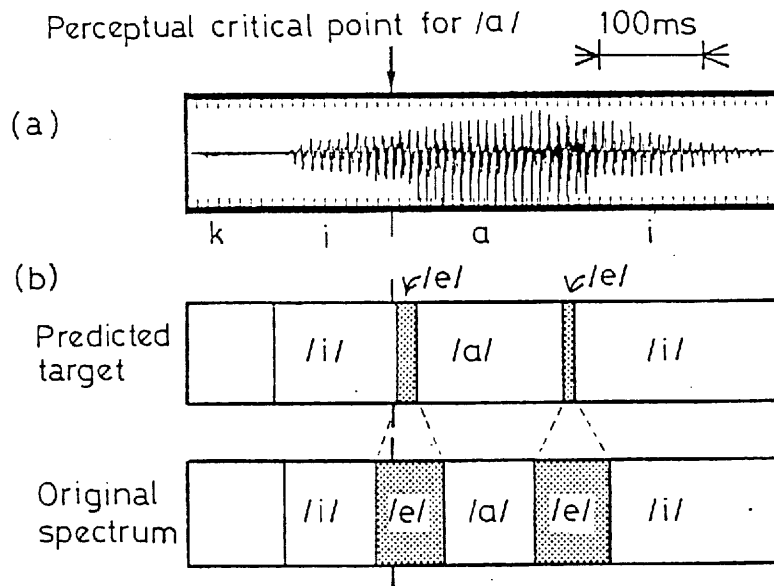


Fig.11 - (a) Speech wave; (b) Time sequences of vowel nearest to the predicted target spectrum or to the original spectrum, at each short period of the Japanese word /kiai/.

important roles than instantaneous spectral features in the human perception process for continuous speech. One of the established methods for this purpose is the Δ cepstrum approach which is widely used in various speech recognition systems. The other method is the spectral-segment-based approach, which has recently been tried. How to define, construct, and train the segments are problems for the future, and should be tackled carefully, using mathematical approaches.

ACKNOWLEDGEMENT

The author wishes to thank the members of the Speech and Acoustics Laboratory and the Honda Research Group of the Basic Research Laboratories for their fruitful discussions.

REFERENCES

- [1] S.Furui: "On the role of spectral transition for speech perception", J.Acoust.Soc.Amer., 80, 4, pp.1016-1025 (1986)
- [2] L.R.Bahl, F.Jelinek and R.L.Mercer: "A maximum likelihood approach to continuous speech recognition", IEEE Trans.Pattern Analysis, Machine Intelligence, PAMI-5, 2, pp.179-190 (1983)
- [3] B.T.Lowerre: "The Harpy speech recognition system", Ph.D.Thesis, Computer Science Dept., Carnegie-Mellon Univ. (1976)
- [4] N.Sugamura and S.Furui: "Large vocabulary word recognition using pseudo-phoneme templates", Trans.IECE, J65-D, 8, pp.1041-1048 (1982)
- [5] K.Shikano: "Spoken word recognition based upon vector quantization of input speech", Trans.Committee on Speech Research, Acoust.Soc.Jap., S83-60 (1982)
- [6] K-C.Pan, F.K.Soong and L.R.Rabiner: "A vector-quantization-based preprocessor for speaker-independent isolated word recognition", IEEE Trans.Acoust.,Speech,Signal Processing, ASSP-33, 3, pp.546-560 (1985)
- [7] S.Furui: "A VQ-based preprocessor using cepstral dynamic features for large vocabulary word recognition", Proc.ICASSP 87, Dallas, 27.2 (1987)
- [8] T.Kohonen, K.Torkkola, M.Shozakai, J.Kangas and O.Venta: "Phonetic typewriter for Finnish and Japanese", Proc.ICASSP 88, New York, S13.10 (1988)
- [9] C-H.Lee, F.K.Soong and B-H.Juang: "A segment model based approach to speech recognition", Proc.ICASSP 88, New York, S11.4 (1988)
- [10] M.Okada and M.Kohda: "Word spotting method based on phonetic segment units", Trans.Committee on Speech Research, Acoust.Soc.Jap., SP87-89 (1987)

- [11] S.Roucos and M.O.Dunham: "A stochastic segment model for phoneme-based continuous speech recognition", Proc.ICASSP 87, Dallas, 3.3 (1987)
- [12] R.Schwartz, Y-L.Chow, O.A.Kimball, S.Roucos, M.Krasner and J.Makhoul: "Context-dependent modeling for acoustic-phonetic recognition of continuous speech", Proc.ICASSP 85, Tampa, 31.3 (1985)
- [13] K-F.Lee and H-W.Hon: "Large-vocabulary speaker-independent continuous speech recognition using HMM", Proc.ICASSP 88, New York, S3.7 (1988)
- [14] S.Sagayama: "Phoneme environment clustering - principle and algorithm", Trans.Committee on Speech Research, Acoust.Soc.Jap., SP87-86 (1987)
- [15] K.Shikano, K-F.Lee and R.Reddy: "Speaker adaptation through vector quantization", Proc.ICASSP 86, Tokyo, 49.5 (1986)
- [16] R.Schwartz, Y-L.Chow and F.Kubala: "Rapid speaker adaptation using a probabilistic spectral mapping", Proc.ICASSP 87, Dallas, 15.3 (1987)
- [17] S.Furui: "Text-independent speaker adaptation based on hierarchical spectral clustering", IEEE Arden House Workshop on Speech Recognition (1988)
- [18] Y.Yamashita and H.Matsumoto: "Speaker adaptation for word recognition based on error vectors in VQ", Trans.Committee on Speech Research, Acoust.Soc.Jap., SP87-118 (1988)
- [19] Y.Shiraki and M.Honda: "Speaker adaptation algorithms for segment vocoder", Trans.Committee on Speech Research, Acoust.Soc.Jap., SP87-67 (1987)
- [20] M.Akagi and S.Furui: "Modeling of vowel target prediction mechanism in speech perception", Trans.IECE, J69-A, 10, pp.1277-1285 (1986)