

論文 / 著書情報
Article / Book Information

Title	Recent Advances in Speech Recognition
Authors	Sadaoki Furui
Citation	Proc. EUROSPEECH 1991, , , pp. 3-10
Pub. date	1991, 9

RECENT ADVANCES IN SPEECH RECOGNITION

Sadaoki Furui

NTT Human Interface Laboratories
Musashino-shi, Tokyo, 180 Japan

ABSTRACT

This paper introduces recent research activities on speech recognition, ranging from acoustic processing to linguistic processing, at NTT (Nippon Telegraph and Telephone Corporation) Laboratories. These include the proposal of hierarchical Δ cepstral parameters and Δ LSP parameters, a new method of utilizing pitch information, automatic speaker adaptation techniques, robust HMM phoneme models, new training algorithms for neural networks, linguistic processing using syntactic and semantic knowledge, implementation of prototype continuous speech recognition systems, and an efficient text-independent speaker recognition algorithm.

I. INTRODUCTION

Speech recognition technology has recently made remarkable progress in various aspects, especially in speaker-independent continuous speech recognition [Mariani, 1989, Furui, 1989a]. This paper discusses recent advances in speech recognition technology, which are closely related to advanced AI, signal processing, statistical modeling and various other technologies. This paper, however, is not intended to be a comprehensive review of speech recognition covering all of the recent technologies. Rather, it is intended to outline major technologies being investigated at NTT Laboratories.

The most important target of the present speech recognition research is to achieve speaker-independent, large-vocabulary, continuous speech recognition systems. In order to bring forth these systems, various kinds of problems must be solved. Figure 1 shows the principal structure of the continuous speech recognition systems, as well as major problems under investigation.

Flow of the recognition process is as follows. First, endpoints of speech periods are detected, typically using the short-time energy level of input speech. Then, the speech wave is converted into a time sequence of feature parameters through spectral analysis.

Cepstral and Δ cepstral coefficients, representing instantaneous and dynamic features of spectra, are widely used. We recently evaluated two new methods: a method of using hierarchical dynamic features (Subsection 2.1) and a method of using LSP (line spectrum pair) and Δ LSP parameters (Subsection 2.2). We have also proposed a new method of using prosodic features (Subsection 2.3), which have hardly been successfully used in speech recognition. The prosodic features are useful for recognizing several words that are phonetically similar and that have different accent patterns.

Although a high speaker-independent recognition accuracy can be obtained by training the recognition system using utterances from a large number of speakers, this method has a performance limitation. It is therefore necessary to introduce techniques of automatically adapting the recognition system to a new speaker (Subsections 3.1, 3.2 and 3.3).

Although word templates or word models are convenient to use in isolated word recognition, units smaller than words are essential for recognizing large-vocabulary, continuous speech. It is also important for so-called co-articulation effects to be considered when creating these units. We have proposed an efficient method of creating context-dependent phoneme units using a clustering technique (Subsection 4.1). We have compared several representations of HMM phoneme models, from the viewpoint of robustness against variations in speaking styles, such as the differences between word-by-word utterances, phrase-by-phrase utterances, and continuous speech (Subsection 4.2).

Although neural-network phoneme modeling has various new possibilities, it has not yet surpassed the performances of the HMM (hidden Markov model)-based methods. New methods of training neural networks have been investigated to stabilize performance (Subsection 4.3). Phoneme recognition results obtained using these techniques and a word dictionary are then combined to recognize words.

The last and the most important stage of the continuous speech recognition systems is the linguistic processing stage, where

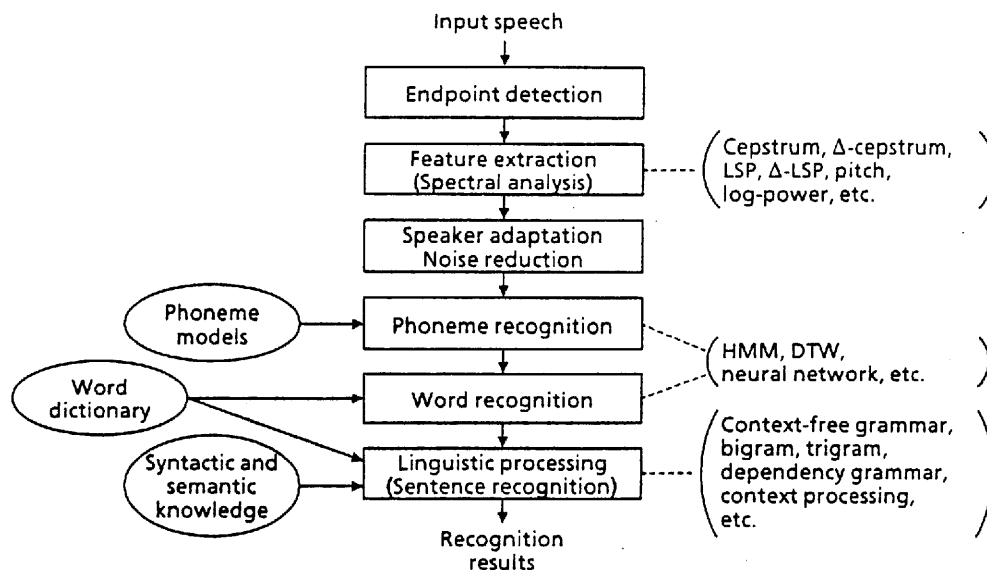


Fig. 1 - Principal structure of continuous speech recognition systems.

syntactic and semantic knowledge are used. Local language modeling, such as transitional network grammar, bigram- or trigram-based grammar modeling, has recently been used in many recognition systems. How to combine these methods with a global syntax such as case grammar is one of the important research subjects (Subsection 5.1).

Applications of continuous speech recognition can be classified into two categories; voice-input word processors (dictation systems), and various information services, such as guidance, reservation, and order receiving services using natural conversational speech. A dictation prototype system and an experimental dictation system, which we recently implemented, will be introduced in Subsections 5.2 and 5.3. We have also implemented a word spotting hardware into a dialog system (Subsection 5.4).

Speaker recognition using speaker-specific information in speech waves is also an important research subject; it is expected to be used in various applications such as security control in the future (Section 6). Section 7 will be devoted to describing future research topics.

II. FEATURE EXTRACTION

2.1 Hierarchical Spectral Dynamics

The effectiveness of a dynamic spectral feature set, namely, that includes $\Delta\text{cepstrum}$ and Δpower , to improve recognition accuracy, especially in speaker-independent recognition, has been confirmed [Furui, 1986], and these parameters are now widely used. The parameters are defined as first-order regression coefficients for short-time sequences of cepstral coefficients and logarithmic energy, respectively. They are usually used in combination with conventional cepstral coefficients and logarithmic energy. The length of the parameter sequence for regression analysis is usually set empirically between 40 ms and 100 ms.

Recently we proposed a recognition method that uses hierarchical spectral dynamic features extracted over multiple time lengths (Fig. 2), and showed the effectiveness of these features in phoneme recognition and isolated word recognition [Furui, 1990a]. Δc_l and Δp_l in Fig. 2 indicate regression coefficients calculated over l frames (7 frames = 56 ms, 21 frames = 168 ms) for cepstrum and log-energy, respectively. The speaker-independent isolated word recognition experiments were performed using a vocabulary of 100 Japanese words.

In this method, input speech is quantized by word-specific codebooks created as subsets of a universal codebook (Fig. 3). When VQ distortion is used for word identification, a low recognition error rate of 3.6% is achieved. When VQ distortion is used for preprocessing, that is, pre-selection of candidate words, the number of candidates for each input utterance is reduced to 1% of the vocabulary, with no increase in the error rate. Phoneme recognition experiments were also performed for the consonants /b/, /d/ and /g/ in a large vocabulary of isolated words uttered by one male speaker. Using the proposed recognition method, the high recognition accuracy of 98-99% was obtained. Since the VQ-distortion method does not require time alignment, it has the advantages of less computation and ease of parallel processing.

2.2 LSP Dynamic Features

LSP frequency representation of a speech signal was introduced as an alternative linear predictive coding (LPC) representation for the purpose of maintaining voice quality at smaller bit rates [Itakura, 1975]. This representation functions in the frequency domain, and various researchers have made use of this representation in various speech applications, such as speech coding, synthesis, and recognition. It has been reported that LSP-based distance measures lead to better recognition performance than conventional distance measures [Paliwal, 1990].

We have introduced distance measures based on the linear

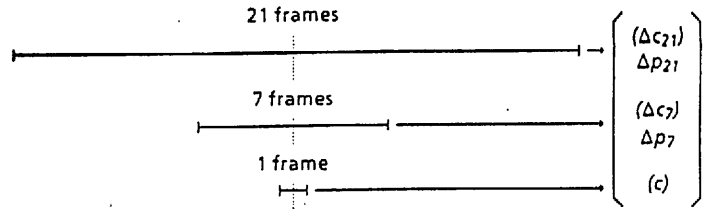


Fig. 2 - Structure of feature vectors for the hierarchical dynamic feature method.

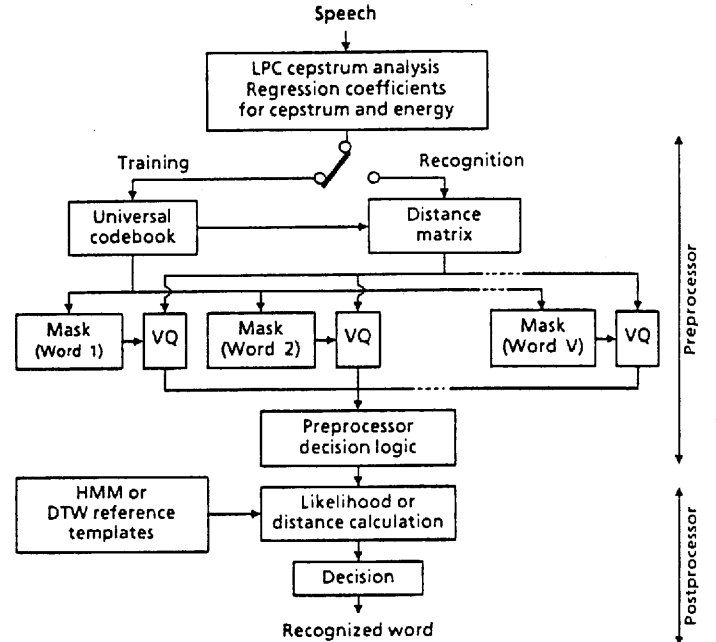


Fig. 3 - Block diagram of word recognizer incorporating a VQ preprocessor and an HMM- or DTW-based postprocessor.

combination of the transitional and the instantaneous LSP frequencies. We have also compared two ways of combining these two sets of parameters;

[Combination in the distance domain]

$$d_{LSP} + d_{\Delta LSP} = \sum_{i=1}^m (x_i - y_i)^2 + w \sum_{i=1}^m (\Delta x_i - \Delta y_i)^2 \quad (1)$$

where, x_i and y_i are the i -th LSP frequency of input speech and reference speech, respectively, and w is a weighting factor set to the inverse of the mean variance.

[Combination in the parameter domain]

$$\begin{aligned} d_{LSP + \Delta LSP} &= \sum_{i=1}^m \{(x_i + w\Delta x_i) - (y_i + w\Delta y_i)\}^2 \\ &= \sum_{i=1}^m \{(x_i - y_i) + w(\Delta x_i - \Delta y_i)\}^2 \end{aligned} \quad (2)$$

Experimental results show that the latter method has a slightly better recognition performance than the former method, and that the performance is better than that obtained using the cepstral and $\Delta\text{cepstral}$ coefficients. Combination in the parameter domain is also advantageous in terms of computation; that is, the

combination can be obtained during speech signal analysis and thus, it does not result in extra computation at the recognition stage [Gurgen et al., 1990].

2.3 On the Use of Prosodic Features in Speech Recognition

Speech wave conveys both spectral information and prosodic information. Prosodic information includes pitch, duration, and power, which are difficult to use in speech recognition. We have recently tried to use pitch information in both isolated word recognition and continuous speech recognition.

Figure 4 shows a block diagram of the word recognition system. The system has two major flows. The left flow is a phonetic recognition part based on HMMs. The other side is for pitch pattern recognition. Pitch pattern of input speech is compared with pitch pattern templates using a DTW technique. Pitch pattern templates are produced by averaging pitch patterns obtained from a set of words having the same accent pattern. A distance measure has been proposed based on a combination of the pitch pattern distance and the phonetic likelihood. Speaker-dependent word recognition experiments were carried out, and it was shown that the proposed measure reduces the recognition error rate from 1.62% to 0.88%, which is roughly a 45% reduction compared with the conventional phonetic likelihood measure [Takahashi et al., 1990].

This method has been extended to continuous speech recognition, in which pitch patterns of Japanese minimal phrases are modeled by HMMs and used for phrase boundary detection. The experimental results show that roughly 70% of the phrase boundaries can be correctly detected by this method [Takahashi et al., 1991].

III. SPEAKER ADAPTATION

3.1 Limits of Speaker-Independent Methods

A number of approaches have been tried in an effort to build speaker-independent recognition systems, typically under HMM-based frameworks. An HMM is a powerful and accurate stochastic model, which can represent speech variations when a sufficiently large amount of training speech is available.

One of the disadvantages of the speaker-independent approach is that it neglects various useful characteristics of the speaker in spite of the fact that they can be learned after recognition of several words or sentences [Furui, 1990b]. If these characteristics can be properly used, the recognition process is expected to be accelerated due to the narrowing of the search space. Another disadvantage is that when the distributions of feature parameters are very broad or multi-modal, such as when male and female voices as well as various dialects are combined, it is difficult to separate phonemes using speaker-independent methods. In these situations, nonsensical matches are probable. For example, the first and the second halves of a word may be matched against the male and female halves of a bimodal distribution, respectively. To cope with these problems, it is essential to introduce speaker-adaptation techniques.

Speaker adaptation or normalization ("speaker adaptation" will indicate both adaptation and normalization hereafter) is a method of automatically adapting reference templates to each new speaker or normalizing (reducing) interspeaker variations in each input speech, based on the transformation rules obtained using a few training words or short sentences. In large vocabulary recognition systems, training with respect to the utterances of all the vocabulary words is too troublesome for the user and consequently unrealistic. Training on a few words or short sentences is, therefore, a practical and realistic solution, although it is less convenient than the speaker-independent systems which have no need for training at all.

Speaker-adaptation methods are generally classified into supervised (text-dependent) methods in which training words or sentences are known, and unsupervised (text-independent)

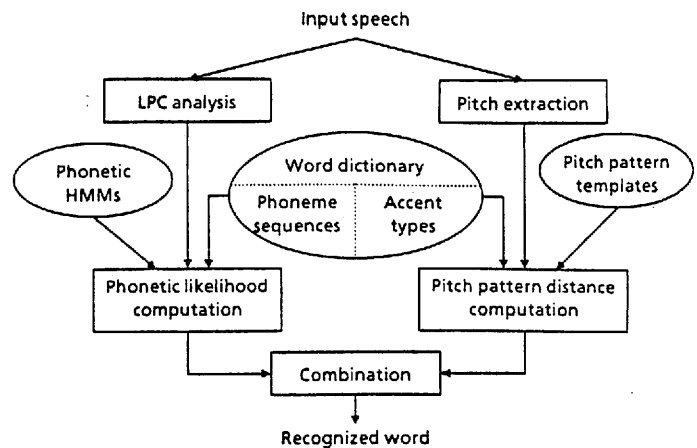


Fig. 4 - Block diagram of word recognizer using spectral and pitch pattern information.

methods in which arbitrary utterances can be used. Both methods can also be classified as offline methods in which training words or sentences must be uttered before the recognition, or as online methods in which utterances for recognition are, at the same time, used for training.

Ideally for users, the system should work as if it were a speaker-independent system which requests no additional training utterances from each speaker. The system should also adapt to the speaker's voice automatically using utterances for recognition. Such a system can be achieved by the unsupervised, online adaptation mechanism.

3.2 Codebook Adaptation/Normalization Method

Speaker-adaptive parameters can be estimated from speaker-independent parameters according to mapping rules. The mapping rules are estimated from the relationships between speaker-independent and speaker-dependent parameters. Within the framework of VQ-based speech recognition, both supervised and unsupervised methods of adapting the speaker-independent codebook to a new speaker or normalizing (adjusting) the input speech to the codebook have been proposed. Each word is represented by an HMM or single/multiple sequences of codebook entries in the word dictionary. Individual variations on how a word is uttered are modeled by the HMM or the multiple code sequences according to these methods. The HMM and code sequences are not changed during the adaptation and are universally used for all speakers.

3.2.1 Supervised Adaptation

For supervised adaptation, the mapping rules are obtained through DTW or a forward-backward algorithm. Figure 5 is a block diagram of a supervised adaptation method using DTW [Shikano et al., 1986]. The utterances of a reference speaker are used to create an initial codebook. These utterances are vector-quantized, that is, converted into sequences of codebook entries. In the training stage, training utterances of a new speaker are converted into code sequences and time-aligned with the same word or sentence uttered by the reference speaker, using the DTW technique. The spectral mapping function between the codebook elements of these two speakers is obtained from alignment functions, that is, the correspondences between the time axes.

Each codebook element is included in various words, and each codebook element of the reference speaker corresponds to various elements of the new speaker. Thus, a histogram of correspondences between codebook elements of the reference speaker and the new speaker, that is, a histogram of co-occurrences of codebook elements, is calculated using the alignment results of all training words or sentences. The mapping function is weighted by the

histogram to find the best correspondence rule.

In the recognition stage, input speech is vector-quantized and mapped (normalized) to the reference speaker's spectrum using the mapping rules at every frame. The similarity between the normalized input speech and each word of the reference speaker is then calculated and used in the recognition decision.

3.2.2 Unsupervised Adaptation

Figure 6 is a block diagram of an unsupervised codebook adaptation method [Furui, 1989b, Furui, 1989c]. The idea of this method is based on an adaptation algorithm for a segment vocoder [Shiraki et al., 1990]. First, an initial codebook and a VQ-indexed word dictionary are prepared. The codebook size was set as 1,024 in the experiment. The initial codebook is produced by clustering the voices of multiple speakers, and commonly serves as the initial condition for each new speaker.

In the adaptation process, a set of spectra from the training utterances of a new speaker and the reference codebook elements are clustered hierarchically in an increasing number of clusters. Using the deviation vectors between centroids of the training spectra clusters and the corresponding codebook clusters, either codebook elements or input frame spectra are shifted so that the corresponding centroids coincide. Continuity between adjacent clusters is maintained by determining the shifting vectors to be the weighted-sum of the deviation vectors of adjacent clusters. Adaptation is thus performed hierarchically from global to local individuality.

Figure 7 illustrates the hierarchical adaptation procedures for shifting the codebook elements from the beginning to the four-cluster stage. The size of the codebook (1,024) is maintained throughout the adaptation process. Using the hierarchical technique, phonetic correspondence between training utterance spectra and codebook elements is roughly maintained.

Several modifications to the adaptation method have also been investigated [Furui, 1989b].

3.3 HMM Adaptation

We recently proposed a new speaker-adaptive speech recognition method using a stochastic speaker classifier [Imamura, 1991].

In the conventional statistical speech recognition method, the main problem was to find the best word sequence W' which maximizes the *a posteriori* probability $P(W|Y)$ for input acoustic string Y . Using Bayes' rule, the *a posteriori* probability $P(W|Y)$ is computed as

$$P(W|Y) = P(W)P(Y|W)/P(Y) \quad (3)$$

where $P(Y)$ is the *a priori* probability of acoustic string Y , $P(W)$ is the *a priori* probability of word sequence W which is given by a language model, and $P(Y|W)$ is the category-conditional probability which is given by an acoustic model such as a sub-word or whole-word HMM.

In our method, the *a posteriori* probability is given by

$$P(W|Y, S) = P(W)P(Y, S|W)/P(Y, S) \quad (4)$$

where $P(Y, S)$ is the *a priori* joint probability of acoustic string Y and speaker individuality S , and $P(Y, S|W)$ is the category-conditional joint probability which is given by a speaker constrained acoustic model. The major problems are; "How to measure the speaker individuality S ?" and "How to compute the $P(Y, S|W)$?"

Figure 8 shows a block diagram of the speaker-adaptive recognition method. In this method, the stochastic speaker classifier is used as the feature extractor for speaker individuality information. The speaker classifier includes several (K) speaker classes represented by speaker Markov models. The speaker

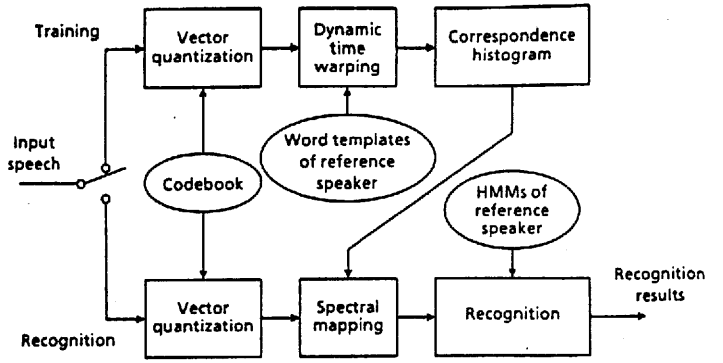


Fig. 5 - Block diagram of supervised speaker adaptation by spectral mapping.

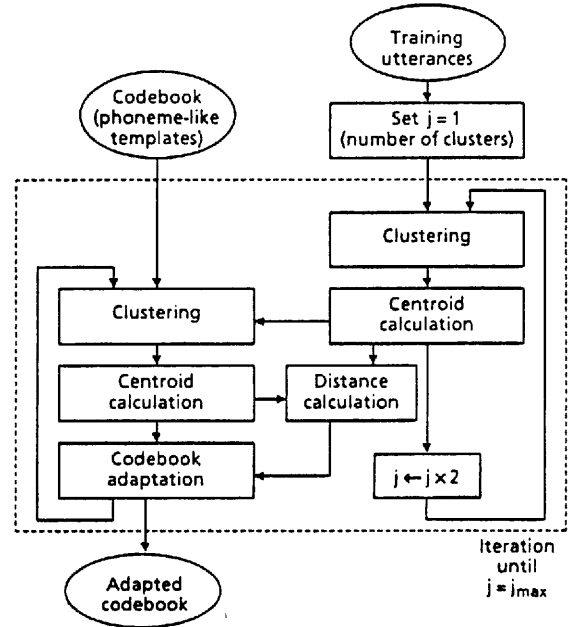


Fig. 6 - Block diagram of unsupervised codebook adaptation.

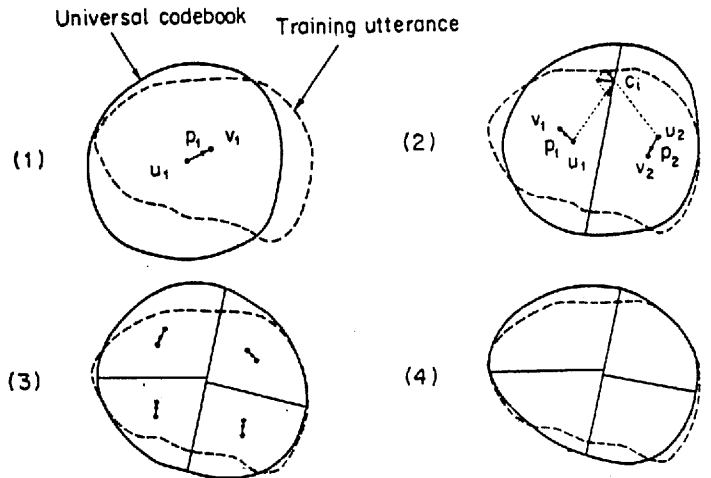
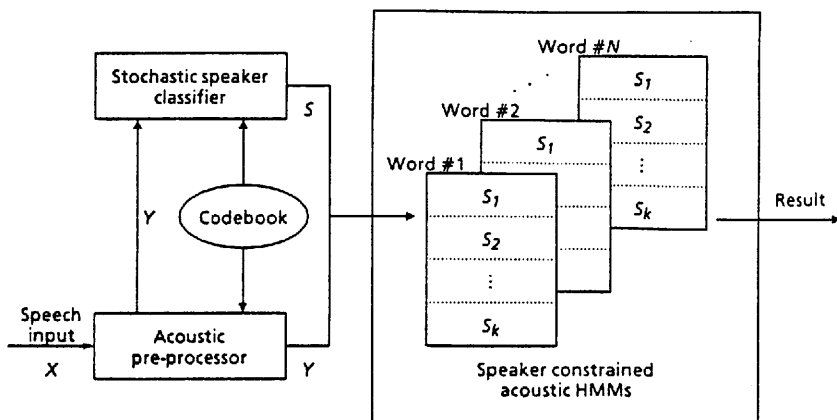


Fig. 7 - Hierarchical codebook adaptation algorithm maintaining continuity between adjacent clusters. v_j : centroid of j -th training spectra cluster, u_j : centroid of corresponding codebook cluster, p_j : deviation vector, and c_i : shifting vector for i -th codebook element.

Fig. 8 - Structure of a speaker-adaptive speech recognition method using a stochastic speaker classifier.



Markov models are estimated by clustering the training speech uttered by many speakers. For each input speech token Y , the classifier computes the category-conditional probability $P(Y|S)$ for each speaker class S . This probability indicates which speaker class (feature sub-space) is most suitable for the input speech. The obtained category-conditional probabilities $P(Y|S)$ are considered to be the quantization results of speaker individuality. In the subsequent word decoding phase, the speaker constrained acoustic HMMs which consider the output of the speaker classifier and the output of the acoustic pre-processor as input symbols are used to compute $P(Y, S|W)$.

Evaluation experiments were performed using a telephone speech database of 50 command words and 10 Japanese digits. Using four 9-state ergodic speaker HMMs estimated from the command words uttered by 116 training speakers, a word recognition accuracy of 98.1% was achieved for the 10 digits uttered by 116 different test speakers. The error rate is half of that obtained by the conventional speaker-independent (pooled training) method.

IV. PHONEME MODELS

4.1 Phoneme Units for Large-Vocabulary Recognition

In large-vocabulary recognition systems, it is impossible to store all words as spectral time sequences. Therefore, it is desirable to prepare phonetic units and to store each word as a sequence of these indices. Since phoneme spectra within continuous speech vary as a result of the influence of preceding and succeeding phonemes, we proposed a method of automatically creating context-dependent phoneme units by means of statistical analysis of a speech database [Sagayama, 1989]. Vocabulary words are represented by a tree-structured concatenation of these units.

4.2 Robustness of HMM Phoneme Models against Speaking Style Variations

For use in real situations, recognition systems must be able to cope with speech variations. The variations can be classified as those due to the speaker and those due to the surrounding environment. The former includes individuality, dialect, fluency, stress, speaking rate, level, pitch, and so on. Variations of phonetic features and transitions between them can be represented efficiently and flexibly in a probabilistic manner by HMMs, which are trained using utterances from many speakers. The HMM method has another advantage in that it uses relatively little computation for recognition, since each word or phoneme is represented by a sequence of a small number of states.

We have examined robustness of six types of phoneme-HMMs against speaking-style variations [Matsuoka et al., 1991]. The six types are VQ- and fuzzy VQ-based discrete HMMs, and single-Gaussian and mixture-Gaussian HMMs with either diagonal or full covariance matrices. Eighteen Japanese-consonant

recognition experiments were performed using isolated word utterances, phrase-by-phrase utterances, and fluently spoken sentence utterances. The mixture-Gaussian HMM with diagonal covariance matrices, the fuzzy VQ-based discrete HMM and the single-Gaussian HMM with full covariance matrices displayed better results than the other three types, when different speaking-style utterances were used in the training and testing.

We have also proposed a new model-adaptation technique that combines multiple models using the deleted interpolation method.

4.3 Fuzzy Training of Neural Net Phoneme Models

Generalization is one of the important issues in using neural networks (NNs) for speech recognition. From this point of view, we have proposed a fuzzy training algorithm, and applied it to phoneme recognition by one-hidden-layer tied-connection NN [Gurgen et al., 1991]. This algorithm is an alternative to the conventional mean square error (MSE) back propagation with hard-decision supervision. The conventional algorithm uses a hard decision criterion (1 and 0 supervisor signals) and suffers from the overlearning problem. In contrast, the proposed training algorithm uses fuzzy-decision supervision. In phoneme sample space, the proposed algorithm takes into account the overlap regions of phoneme boundaries.

The supervisor signal of each sample of training data is determined by the grade of membership using neighboring samples including the original sample. The grade of membership for each component F_k of the supervisor signal is

$$F_k = \sum_{j \in C_k} \exp(-d_j) / \sum_{j=1}^N \exp(-d_j) \quad (5)$$

where C_k is the k -th phoneme class, d is the Euclidean distance, and N is the number of nearest neighbors. The fuzzy effect of N -nearest neighbors on each training sample is computed and the NN is trained with these training samples using the MSE criterion. This method therefore softens the hard decision criterion and reduces the overlearning problem. Initiation of the fuzzy training is done with the conventional hard-decision back propagation method to achieve a good initial weight set.

The proposed algorithm provides better generalization for different speaking styles (isolated word, phrase, and continuous speech) than the conventional back propagation algorithm.

V. CONTINUOUS SPEECH RECOGNITION SYSTEMS

5.1 System Using Dependency Analysis

In our continuous speech recognition system (Fig. 9) [Matsunaga et al., 1990], input speech is represented by time

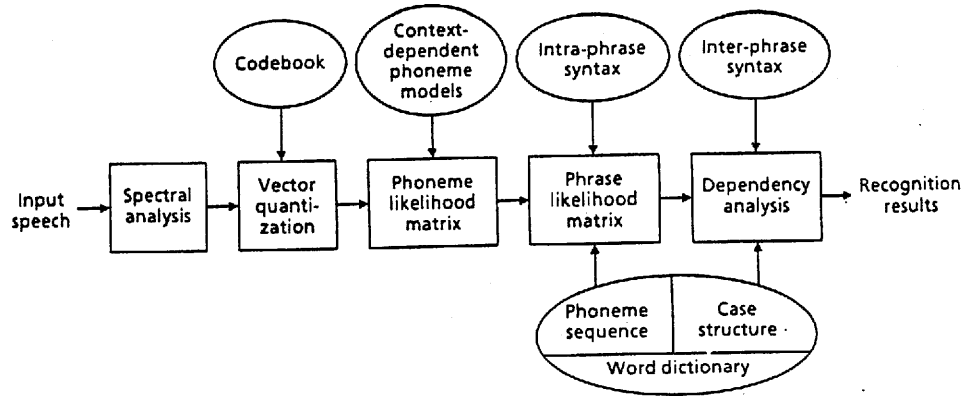


Fig. 9 - Structure of continuous speech recognition system based on two-level grammar approach.

sequences of cepstral and Δ cepstral coefficients. The time sequences are then converted into code sequences using a vector quantization technique. The vector code sequences are compared with the context-dependent phoneme units and likelihoods for the occurrence of various phonemes are stored in a likelihood matrix. Then likelihoods for the occurrence of phrases are calculated using the phoneme likelihood matrix, a Japanese word dictionary, and a transition network grammar for constructing Japanese phrases.

Finally, the sentence is recognized as a sequence of most likely phrases by considering inter-phrase syntactic and semantic likelihoods calculated on the basis of an inter-phrase dependency analysis. A joint likelihood, combining acoustic, syntactic, and semantic likelihoods derived from acoustic processing and linguistic processing, is maximized to obtain the optimum solution. The procedure takes into account the redundancy of speech and the large freedom of phrase order which is a characteristic of the Japanese language.

The system was evaluated using a 71 sentence speech (418 phrases) uttered by two speakers, one male and one female. Two training sets were used: 216 phonetically balanced words, and a combination of 216 words and 29 sentences. The word dictionary has 360 entries and perplexity of the phrase syntax is 40. Certainty factors of dependency relationships were empirically determined through the analysis of technical literature. Table 1 shows the experimental results. The dependency parser increased the average phrase recognition rate from 69.2% to 83.1%, when the

Table 1 - Sentence speech recognition results

HMM training data		216 words		216 words + 29 sentences	
Dependency analysis		no	yes	no	yes
Phrase recognition rate (%)	male	45.5	55.3	63.6	78.9
	female	54.5	75.4	74.8	87.3
	average	50.6	65.3	69.2	83.1

training set of 216 words and 29 sentences was used. These results show the effectiveness of the semantic dependency analysis.

5.2 Japanese Dictation System for Medical Doctors

We are now developing a Japanese dictation system, voice-input word processor, for medical doctors using continuous speech recognition technology [Tsuboi et al., 1990]. The system is speaker-dependent and recognizes continuous phrasal speech.

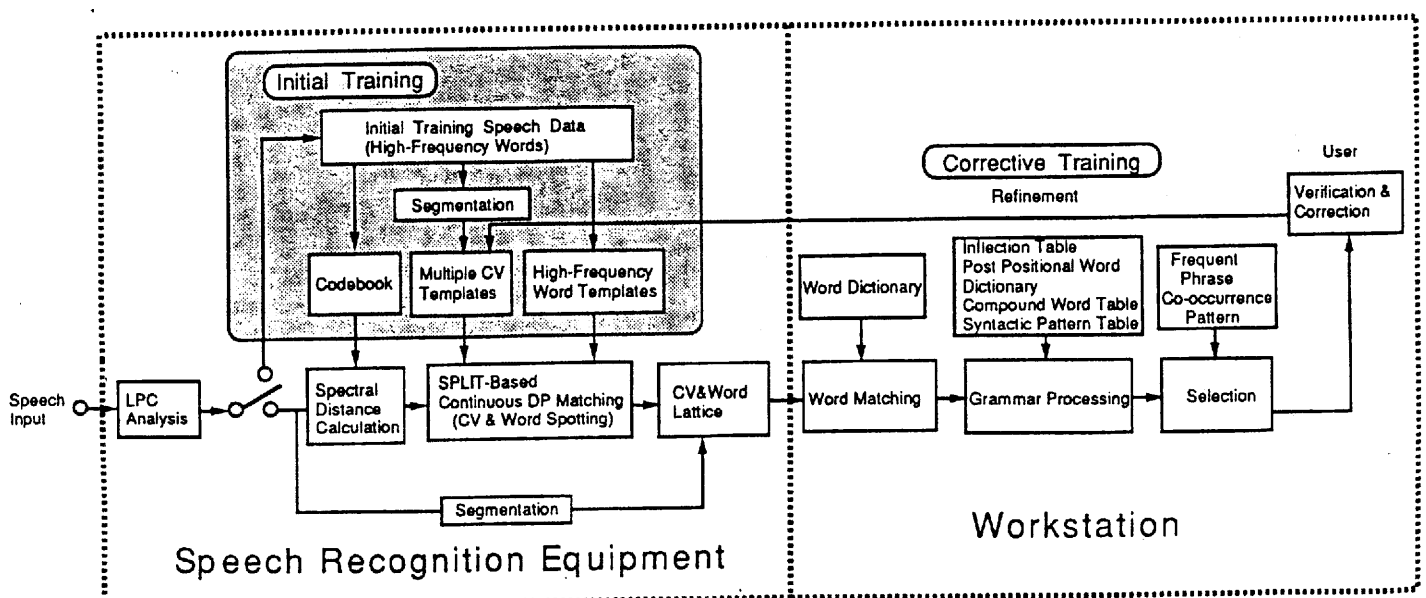


Fig. 10 - Japanese dictation system for medical doctors using continuous speech recognition.

Figure 10 illustrates the structure of the system, which consists of acoustic processing and linguistic processing parts. These two parts are implemented by a speech recognition equipment and a workstation, respectively. In the former part, a CV (consonant-vowel) syllable lattice and a word lattice are generated from input speech using a VQ-based continuous DP-matching technique. These lattices are converted into written Japanese form using a word dictionary and grammar information. The CV templates are automatically adapted to each speaker using the recognition results. Japanese conversion accuracies at this moment for X-ray CT scanning reports are roughly 80% for normal reports, and 65% for abnormal reports.

5.3 Dictation System Using Phoneme Source Modeling

We are investigating a phonetic typewriter that utilizes the underlying syntactic and statistical structure of phoneme and character sequences [Yamada et al., 1991]. A schematic diagram of the system, which consists of an HMM-based acoustic processing part and phoneme source modeling, is shown in Fig. 11. A syllable trigram approach to language source modeling is effective and promising for the Japanese language, since Kana (Japanese syllabary alphabet) roughly corresponds to consonant-vowel (CV) syllables.

For our phonetic typewriter, a general Japanese syllable sequence structure is written using context-free rewriting rules, and this structure is precompiled into the form of an LR table with syllable trigram probabilities. These syllable trigram probabilities are obtained by using a large text database. The predictive LR parser predicts possible phoneme sequences from left to right according to the general Japanese phoneme sequence syntax. The parser calculates phoneme sequence probabilities based on syllable trigram and HMM probabilities, and the system recognizes speech as the phoneme sequence with the highest probability.

The phonetic typewriter has been tested using 279 phrases uttered by one male speaker, and the syllable source model has achieved a 94.9% phoneme recognition rate with the test-set phoneme perplexity of 3.9. Without the syllable trigram, the phoneme recognition rate was only 73.2%.

A trigram model based on character (Kana and Chinese character) sequences in usual Japanese sentences has also been studied. It has been indicated that the character trigram model can significantly reduce the phoneme perplexity, compared with the syllable trigram model.

5.4 Word Spotting System

One of the important issues regarding man-machine interactive systems using speech recognizers is how to cope with noise and non-vocabulary words. As a solution to this problem, we have implemented an HMM-based speaker-independent word spotting system using Transputers [Imamura et al., 1990]. Candidates of word end-points and corresponding likelihood scores are computed with the continuous Viterbi decoding algorithm. To prune unreasonable candidates, three new methods have been introduced to duration control, threshold logic (for the likelihood scores), and local peak detection. An efficient parallel processing scheme for word spotting is carried out by using a tree structure of Transputers. In each frame period, the spectral feature vector from the speech analyzer is broadcasted from a root Transputer to eight node Transputers. With this structure, 72 words can be processed within a 12 ms frame period.

Word spotting experiments were conducted using 10 Japanese digits as keyword vocabulary. The system was trained using a database of keywords uttered isolatedly by 107 male speakers. A database of 100 short sentences, which were uttered by 10 new male speakers and recorded over noisy telephone networks, were used for testing. Each sentence included one keyword. Experimental results show that when the pruning is controlled to set the false alarm rate at 3%, a word detection accuracy of 97% can be obtained.

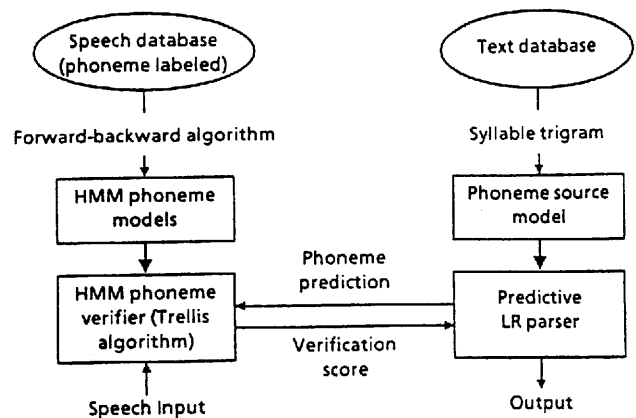


Fig. 11 - Schematic diagram of phonetic typewriter based on the HMM-LR method.

VI. SPEAKER IDENTIFICATION AND VERIFICATION

We have proposed a VQ-based text-independent speaker recognition method robust against text-to-text and inter-session variations [Matsui et al., 1991]. Figure 12 shows the principal structure of the system, incorporating three key techniques to cope with these spectral variations. First, either an ergodic HMM or a V/UV (voiced/unvoiced) decision is used to classify input speech into broad phonetic classes. The figure shows the case where HMM-based classification is used. Second, a new distance measure, Distortion-Intersection Measure (DIM), is introduced for calculating VQ distortion of input speech against speaker-dependent codebooks. DIM is characterized by selective matching using only a stable subset of test speech in VQ distortion calculation. Third, a new feature normalization method, Talker Variability Normalization (TVN), is introduced. TVN normalizes parameter variations taking both inter- and intra-speaker variability into consideration. TVN emphasizes feature parameters that have a relatively large inter-to-intra-speaker variation ratio.

The system was tested using utterances of nine speakers recorded at four sessions over three years. Speaker-dependent codebooks were made at every session, and the utterances recorded at different sessions were tested against these codebooks. When a 2-state ergodic HMM was used for broad phonetic classification, cepstral and Δ cepstral coefficients were used as feature parameters, and an average speaker identification accuracy of 98.5% was obtained. When V/UV decision was used for the classification, cepstral and Δ cepstral coefficients were extracted in unvoiced periods of speech, and pitch and Δ pitch frequencies were additionally extracted in voiced periods. In the latter case, a speaker identification accuracy of 99.0% was achieved.

VII. FUTURE PROBLEMS

Speech recognition technology is expected to play important roles in future communication and information services. To achieve these new services, it is important to consider the following questions from the viewpoint of human-machine interface. "What are the desirable forms of human-machine interface?" "What kinds of systems are really comfortable for users?" "What are the conditions necessary to achieve a system which is helpful for intelligent human activities?"

From a technological point of view, future problems include topics related to speech individuality, robust and proper statistical modeling and new technologies such as sophisticated neural networks. The first topic is how to extract, process and normalize speech individuality. The second involves the problem of how to

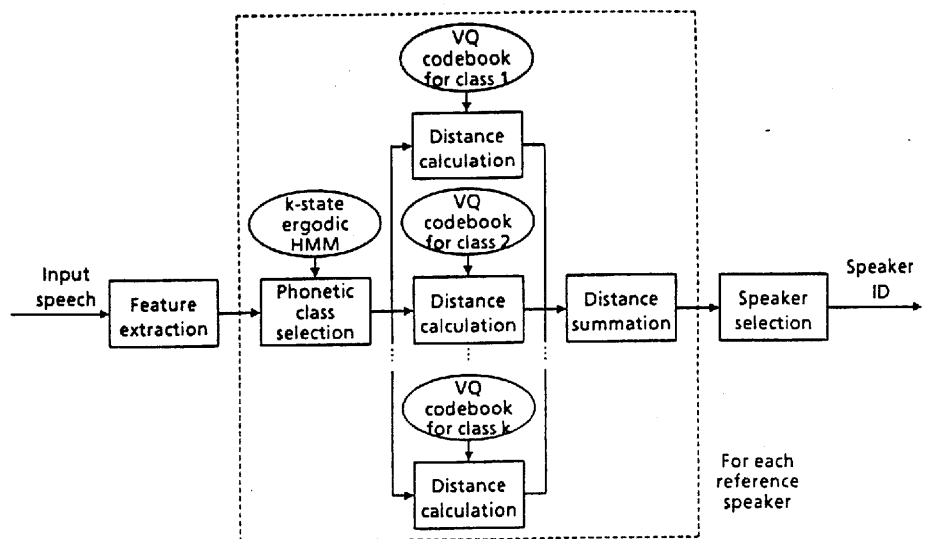


Fig. 12 - Block diagram of a text-independent speaker identification system using HMM-based broad phonetic classification.

produce robust statistical models based on restricted training data. Although HMM technology has brought speech recognition system performance to new high levels for a variety of applications, there still remain some fundamental problems. The third includes "How can we obtain advanced knowledge engineering technology, including knowledge acquisition and learning algorithms, which can handle human common sense?" and "How can we improve the neural network technology and combine it with conventional methods?"

ACKNOWLEDGMENTS

A number of people at NTT Laboratories have contributed to the works described in this paper. The author wishes to thank Dr. Kiyohiro Shikano, head of the Speech Information Processing Group, Dr. Hirokazu Sato, head of the Speech Processing Systems Group, and all staff members for their fruitful discussions and continuous research and development efforts.

REFERENCES

- Furui, S. (1986): "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-34, pp.52-59
- Furui, S. (1989a): "Digital Speech Processing, Synthesis, and Recognition," Marcel Dekker, New York
- Furui, S. (1989b): "Unsupervised speaker adaptation method based on hierarchical spectral clustering," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, Scotland, S6.9
- Furui, S. (1989c): "Unsupervised speaker adaptation based on hierarchical spectral clustering," *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-37, pp.1923-1930
- Furui, S. (1990a): "On the use of hierarchical spectral dynamics in speech recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, S15a.10
- Furui, S. (1990b): "Speaker-dependent-feature extraction, recognition and processing techniques," *Proceedings of the ESCA Tutorial and Research Workshop on Speaker Characterization in Speech Technology*, pp. 10-27
- Gurgen, F., Sagayama, S. and Furui, S. (1990): "Line spectrum pair frequency-based distance measures for speech recognition," *Proc. Int. Conf. Spoken Language Processing*, Kobe, 13.1
- Gurgen, F., Aikawa, K. and Shikano, K. (1991): "The improvement of phoneme recognition performance of a neural network using fuzzy training," *Proc. SYNAPSE'91*, Osaka
- Imamura, A. and Suzuki, Y. (1990): "Speaker-independent word spotting and a transputer-based implementation," *Proc. Int. Conf. Spoken Language Processing*, Kobe, 13.5
- Imamura, A. (1991): "Speaker-adaptive HMM-based speech recognition with a stochastic speaker classifier," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, 58.S13.3
- Itakura, F. (1975): "Line spectrum representation of linear predictor coefficients," *J. Acoust. Soc. Am.*, 57, S35(A)
- Mariani, J. (1989): "Recent advances in speech processing," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, Scotland, S9.1
- Matsui, T. and Furui, S. (1991): "A text-independent speaker recognition method robust against utterance variations" *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, 57.S6.3
- Matsunaga, S., Sagayama, S., Homma, S. and Furui, S. (1990): "A continuous speech recognition system based on a two-level grammar approach," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, S11.7
- Matsuoka, T. and Shikano, K. (1991): "Robust HMM phoneme modeling for different speaking styles," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, 10.S5.4
- Paliwal, K. K. (1990): "A study of LSF representation for speaker-dependent and speaker-independent HMM-based speech recognition systems," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, S15a.13
- Sagayama, S. (1989): "Phoneme environment clustering for speech recognition," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Glasgow, Scotland, S8.3
- Shikano, K., Lee, K.-F. and Reddy, R. (1986): "Speaker adaptation through vector quantization," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tokyo, 49.5
- Shiraki, Y. and Honda, M. (1990): "Speaker adaptation algorithms based on piece-wise moving adaptive segment quantization method," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Albuquerque, S12.5
- Takahashi, S., Matsunaga, S. and Sagayama, S. (1990): "Isolated word recognition using pitch pattern information," *Proc. Int. Conf. Spoken Language Processing*, Kobe, 13.9
- Takahashi, S. and Matsunaga, S. and Shikano, K. (1991): "Accent phrase boundary detection in continuous speech using hidden Markov models," *Proc. Spring Meeting of Acoust. Soc. Jap.*, 2-5-13 (in Japanese)
- Tsuboi, T. and Sugamura, N. (1990): "A prototype for a speech-to-text transcription system," *Proc. Int. Conf. Spoken Language Processing*, Kobe, 20.8
- Yamada, T., Hanazawa, T., Kawabata, T., Matsunaga, S. and Shikano, K. (1991): "Phonetic typewriter based on phoneme source modeling," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Toronto, 56.S3.4