Speech Communication 10 (1991) 505–520
North-Holland

Speaker-dependent-feature extraction, recognition and processing techniques

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

Received 8 March 1991

Abstract. This paper discusses recent advances in and perspectives of research on speaker-dependent-feature extraction from speech waves, automatic speaker identification and verification, speaker adaptation in speech recognition, and voice conversion techniques. Speaker-dependent information exists both in the spectral envelope and in the supra-segmental features of speech. This individual information can be further classified into temporal and dynamic features. Speaker identification/verification methods can be divided into text-dependent and text-independent methods. Although text-dependent speaker verification techniques have almost reached the level suitable for practical implementation, text-independent techniques are still in the fundamental research stage. Both supervised and unsupervised speaker adaptation algorithms for speech recognition have recently been proposed, and remarkable progress has been achieved in this field. Improving synthesized speech quality by adding natural characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another, are as yet little exploited research fields to be studied in the near future. Research on speaker-dependent information is one of the most important future directions for achieving advanced speech information processing systems.

Zusammenfassung. Dieser Beitrag diskutiert Perspektiven der Forschung im Bereich der Bestimmung der Sprechermerkmale im Sprachsignal, der automatischen Identifizierung und Erkennung von Sprechern, der Sprecheranpassung in automatischer Spracherkennung, sowie der Techniken der Stimmumwandlung. Sprecherabhängige Information existiert in der spektralen Hüllkurve und in den suprasegmentalen Merkmalen von Sprache. Diese spezifische Information kann weiterhin unterteilt werden in Zeitmerkmale und dynamische Merkmale. Methoden der Sprecheridentifizierung oder Überprüfung können unterteilt werden in textabhängige oder textunabhängige Methoden. Obwohl textabhängige Methoden der Sprecherüberprüfung fast ein Niveau erreicht haben welches praktische Anwendungen erlaubt, sind textunabhängige Methoden noch im Stadium der reinen Forschung. Sowohl überwachte als unüberwachte Algorithmen zur Sprecheranpassung in der automatischen Spracherkennung sind kürzlich vorgestellt worden und bemerkenswerte Fortschritte sind auf diesem Gebiet erzielt worden. Die Verbesserung der Qualität von synthetischer Sprache durch Zuhilfenahme der natürlichen Charakteristiken der Individualität der Stimme, sowie der Übergang der synthetischen Stimmindividualität von einem Sprecher zum anderen sind noch ungenügend erforschte Gebiete welche in naher Zukunft studiert werden. Das Studium der sprecherabhängigen Information ist eine der wichtigsten zukünftigen Richtungen von denen die Vollendung von Informationsaufbereitungssystemen welche auf Sprache beruhen abhängt.

Résumé. Dans cet article, des développements récents concernant l'extraction, à partir de l'onde de parole, des indices dépendants du locuteur, l'identification et la vérification automatiques du locuteur, l'adaptation au locuteur en reconnaissance automatique de la parole et les techniques de conversion de voix sont discutés. L'information concernant le locuteur se trouve à la fois dans l'enveloppe spectrale et dans les traits prosodiques de la parole. Cette information peut de plus être classée en traits temporels et traits dynamiques. Les méthodes de vérification/identification du locuteur peuvent être divisées en méthodes dépendantes du texte et méthodes indépendantes du texte. Bien que les techniques de vérification du locuteur dépendantes du texte aient presque atteint le niveau de développement approprié pour l'implémentation pratique, les techniques indépendantes du texte en sont toujours au stade de la recherche fondamentale. En reconnaissance de parole, des algorithmes d'adaptation au locuteur supervisés et non supervisés ont récemment été proposés, et des progrès remarquables on été réalisés dans ce domaine. L'amélioration de la qualité de la parole synthétique par l'ajout de caractéristiques vocales individuelles et la conversion de l'individualité vocale synthétique d'un locuteur à l'autre sont des sujets de recherche peu exploités actuellement qui devraient être étudiés dans un proche avenir. La recherche sur l'information dépendante du locuteur constitue l'une des plus importantes directions à suivre pour réaliser des systèmes avancés de traitement de l'information dans le domaine de la parole.

Keywords. Speaker-dependent features, speaker recognition, speaker adaptation, voice conversion.

1. Introduction

Speaker-dependent information plays an important role in our daily life in addition to linguistic (phonetic) information. For example, when we talk with our friends over the telephone, we can easily recognize the other party based on his/her voice without hearing the name. The human capability of distinguishing individual voices also plays a central role in the "cocktail party effect", supplementing the auditory binaural effect. Using these capabilities, we can pick up utterances of a specified speaker from the surrounding noise including other speakers' voices.

This paper presents recent progress in and the future direction of research on extraction of speaker-dependent features from speech waves, automatic speaker recognition, speaker adaptation techniques using these features, and voice conversion (Furui, 1989a, 1990).

2. Speaker-dependent information in speech waves

Speaker-dependent information exists both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech. This individual information can be further classified into temporal and dynamic features. It arises both from hereditary individual differences in articulatory organs, such as the length of the vocal tract, the size of the nasal cavity, and vocal cord characteristics, and from acquired differences in manner of speaking, such as dialect and accent. Hereditary differences (anatomical differences) appear as variations in formant frequencies, bandwidth, mean fundamental frequency, inclination of overall spectrum pattern. Acquired differences (speaker-dependent speaking habits) appear as pitch, speech rate and loudness. They are represented by differences in time functions of fundamental frequency, formant frequencies and word duration. These variations are combined together in the speech waves, making it difficult to separately extract these two types of individual information. Therefore, feature parameters containing both types of information are usually used in individual feature extraction.

The average characteristics of variations in vowel formant frequencies due to gender and age can be modeled by the shift in the logarithmic frequency scale maintaining the relative location of the vowels. However, individual characteristics are extremely complicated. And it has also been observed that mutual effects exist between phonetic and individual information, specifically the differences in individual information depending on the phonemes produced (Furui, 1986). This means that text-independent individual information is hard to extract.

3. Classification of speaker recognition methods

Speaker recognition can be principally divided into speaker verification and speaker identification. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is applicable to various kinds of services which involve the use of voice as the key to confirming the identity claim of a speaker. These services include banking transactions using a telephone network, database access services, security control for confidential information areas, and so on. Speaker identification can be used in criminal investigations to determine which of the suspected persons produced the voice recorded at the scene of the crime. Speaker verification has generally larger application areas than speaker identification.

The difficulty of speaker identification increases with the size of the candidate population. On the other hand, the difficulty of speaker verification does not depend on the population because it involves only a binary decision of acceptance or rejection (Furui, 1989a).

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to issue a predetermined utterance, whereas the latter do not rely on a specific text being spoken. Although text-dependent speaker verification techniques

have almost reached the practical level, text-independent techniques are still in the fundamental research stage with various techniques being investigated.

One of the most difficult problems in speaker recognition is that intersession variability (variability over time) of speech waves and spectra for a given speaker significantly affects recognition accuracy. The spectral equalization (normalization) technique has been confirmed to be effective in reducing long-term spectral variation (Furui, 1974). It is also essential to use physical features which are stable and not easily mimicked or affected by transmission characteristics.

4. Examples of text-dependent speaker recognition methods

4.1. DTW-based method

Figure 1 is a block diagram of a typical DTW (dynamic time warping)-based text-dependent speaker verification system (Furui, 1981). Initially, 10 LPC cepstral coefficients are extracted every 10 ms from a short sentence of speech. These cepstral coefficients are then averaged over the duration of the entire utterance. Spectral equalization is performed by subtracting the averaged values from the cepstral coefficients of every frame to compensate for transmission distortion and intraspeaker variability. In addition to the normalized cepstral coefficients, first- and second-order derivatives of the time functions of cepstral coefficients are extracted every 10 ms to represent spectral dynamics. The time function of the set of parameters is brought into time registration with the reference template in order to calculate the distance between them. The overall distance is then compared with a threshold for the verification decision.

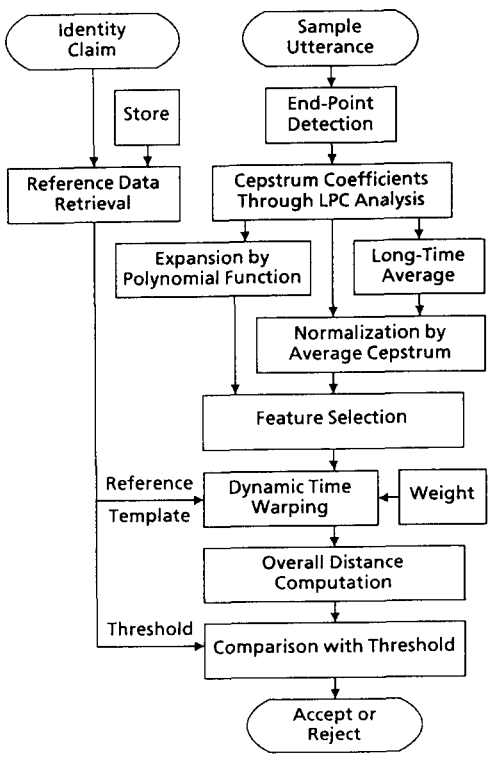

Fig. 1. Block diagram of a DTW-based text-dependent speaker verification system using instantaneous and transitional cepstra.

An online experiment performed over a period of six months, using dialed-up telephone speech uttered by 60 male and 60 female speakers, produced an average recognition accuracy of 97%.

4.2. HMM-based method

HMM (hidden Markov model)-based methods have recently been tried in an effort to improve the recognition performance (Zheng and Yuan, 1988; Naik et al., 1989). Since the HMM has the capability of efficiently modeling statistical variation of spectral features, it achieves significantly better recognition accuracies than do the DTW-based methods.

A speaker verification system based on characterizing the utterances as sequences of subword units represented by HMMs has been introduced and tested (Rosenberg et al., 1990a). Two types of subword units, phone-like units (PLUs) and acoustic segment units (ASUs), have been studied. PLUs are based on phonetic transcriptions of spoken utterances and ASUs are extracted directly from the acoustic signal without use of any linguistic knowledge. Verification performance has been evaluated on a 100-speaker database of 20,000 isolated digit utterances. The results show only small differences in performance between PLU- and ASU-based representations. Overall, the verification equal-error rate

is approximately 7 to 8% for 1-digit test utterances and 1% or less for 7-digit test utterances.

5. Examples of text-independent speaker recognition methods

5.1. Average-spectrum-based method

One of the typical text-independent methods is that using the longtime-averaged spectrum (Furui et al., 1972). In this method, the phoneme effects in speech spectra are removed by averaging the spectra extracted in every short period over the entire utterance, and the effect of long-term spectral variation is reduced by introducing a weighted cepstral distance measure.

5.2. VQ-based method

An extension of the previous method is one using various speaker-specific spectral patterns instead of only the averaged spectrum (Li and Wrench, 1983). The speaker-specific patterns are produced by clustering the spectral distribution of each reference speaker, and are stored as elements in a codebook. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker, and the distribution of the quantization error is calculated over the entire speech interval. The distribution parameters obtained using each reference codebook are examined to make the recognition decision.

An experiment was performed in which the speech signal was analyzed by the LPC technique, the codebook size for each speaker was 40, the number of registered speakers was 11, and the training utterances were 100 s long for each speaker. The results show speaker identification accuracies of 96%, 87% and 79% with 10 s, 5 s and 3 s test utterances.

A method using two VQ codebooks, containing instantaneous and transitional spectral representations, has also been developed (Soong and Rosenberg, 1988). Figure 2 shows a block diagram of the recognition system. Spectral distances, that is, quantization errors from test vectors to the two VQ codebooks, are optimally combined for making the final recognition decision. The experimental results show that since the instantaneous and transitional representations are relatively uncorrelated, they provide complementary information for speaker recognition. They also show that the transitional representations and performance are relatively resistant to simple variations of the transmission channel.

Figure 3 shows a method using a single codebook for long feature vectors consisting of instantaneous and transitional features calculated for both cepstral and vocal source characteristics (Matsui and Furui, 1990). The fundamental frequency and its time derivative are used as source characteristics. Since the fundamental frequency

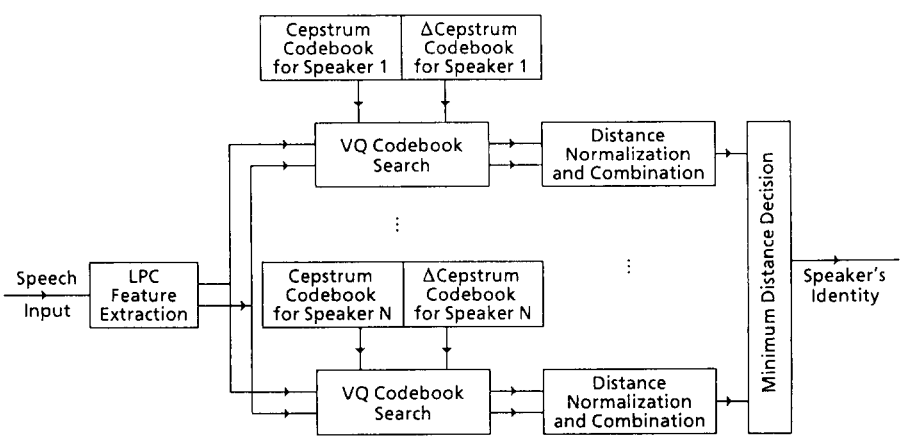

Fig. 2. Block diagram of a VQ-based text-independent speaker identification system.

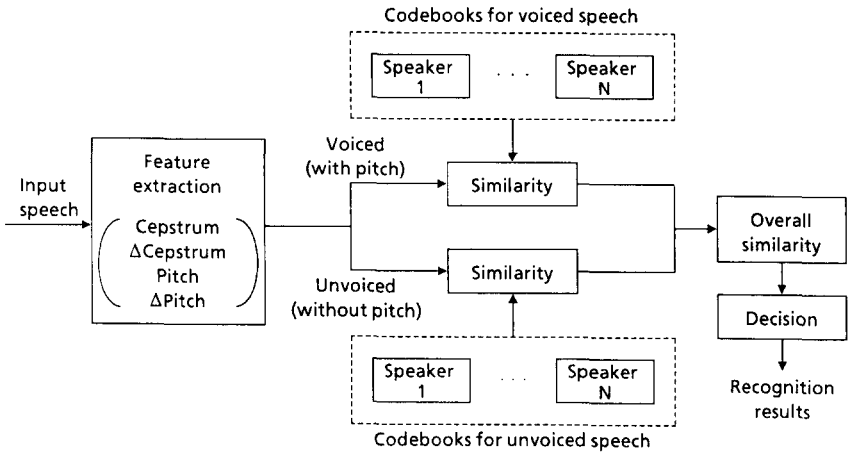

Fig. 3. Block diagram of a VQ-based text-independent speaker recognition system using codebooks representing cepstral and pitch characteristics.

cannot be extracted from unvoiced speech, there are two separate codebooks for voiced and unvoiced speech for each speaker. A new distance measure has also been introduced to take into account the intra- and inter-speaker variability and to deal with the outlier problem of the distribution of feature vectors. The outlier vectors correspond to the intersession spectral variation and the difference between training texts and testing utterances. Experimental results confirm high recognition accuracies even when codebooks for each speaker are made using training utterances recorded at only one session and the time difference between training and testing is more than three months.

In contrast with the memoryless VQ-based

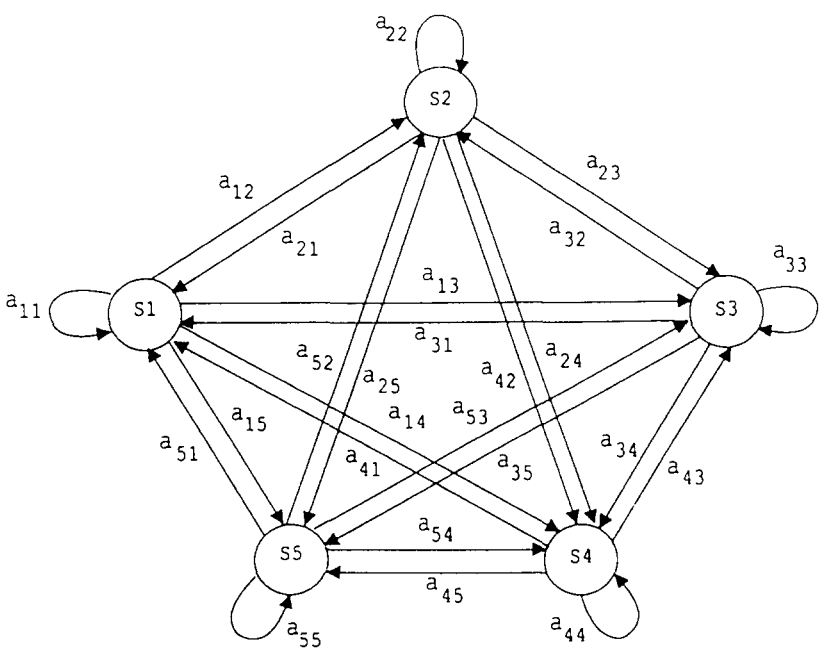

Fig. 4. A five-state ergodic HMM for text-independent speaker verification.

method, non-memoryless source coding algorithms have also been studied (Juang and Soong, 1990). This study examined the effects of source variations, including speaking inconsistency and channel mismatch in source coder designs. It was found that incorporating memory into source coders generally enhances speaker recognition accuracy, though more improvement is possible by including potential source variations in the coder design/training.

5.3. HMM-based method

On a longtime-scale, temporal variation in speech signal parameters can be represented by stochastic Markovian transitions between states. The five state ergodic HMM shown in Figure 4 was used in speaker recognition (Poritz, 1982). Each speaker was represented by a speaker-specific five-state HMM, and good discrimination was achieved among five talkers with 40 s of training data per talker.

This method has recently been improved (Savic and Gupta, 1990). It has been found that different classes of phonemes are not equally effective in discriminating between speakers; therefore, speech segments are separately classified into one of the broad phonetic categories corresponding to the HMM states. Figure 5 shows a

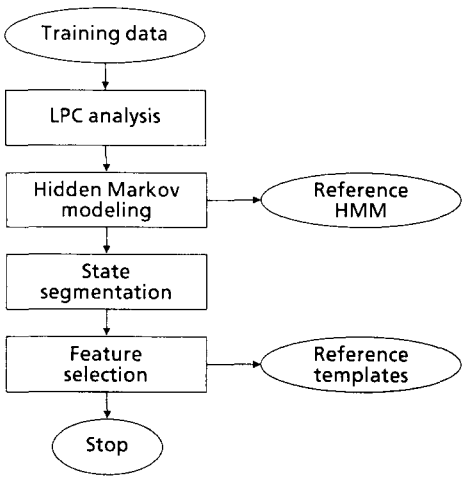

Fig. 5. Functional block diagram of the training process in the HMM-based method.

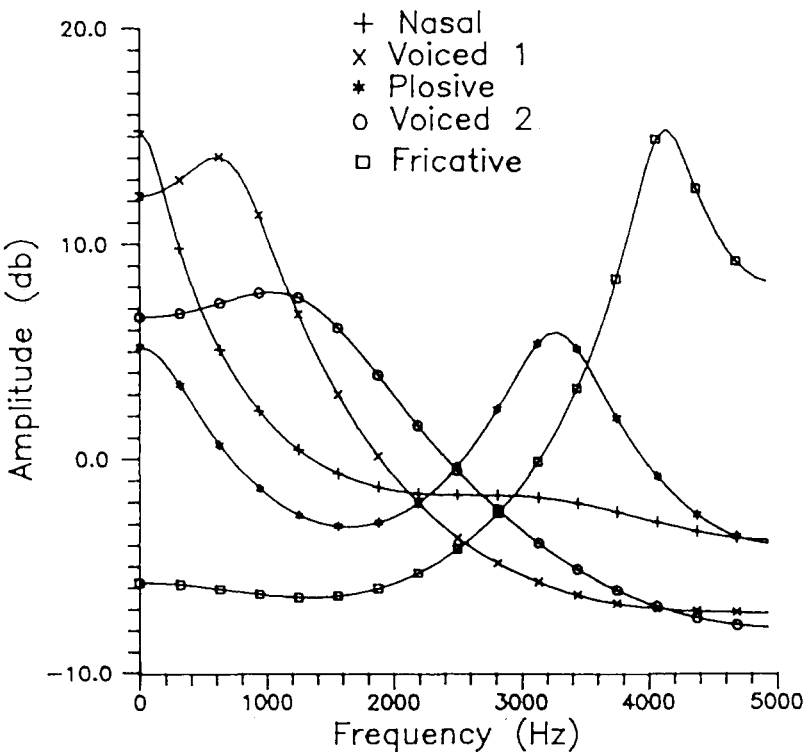

Fig. 6. Typical magnitude spectra corresponding to the five states of HMM.

block diagram of the training process and Figure 6 indicates typical magnitude spectra corresponding to the five states of the HMM. As shown in this figure, the five states correspond to nasal, plosive, fricative and two voiced sounds. Figure 7 shows a block diagram of the verification process. A weighted linear combination of scores for individual categories is used for the final verification decision. Experimental results show that verification accuracy can be considerably improved by this category-dependent weighted linear combination method.

The performance of the speaker recognition method using codebooks representing both cepstral and pitch characteristics, described in Section 5.2, has been improved by introducing an ergodic HMM for broad phonetic categorization (Matsui and Furui, 1991). The broad phonetic categorization can also be implemented by a speaker-specific hierarchical classifier instead of an HMM, and the effectiveness of this approach has been confirmed (Eatock and Mason, 1990).

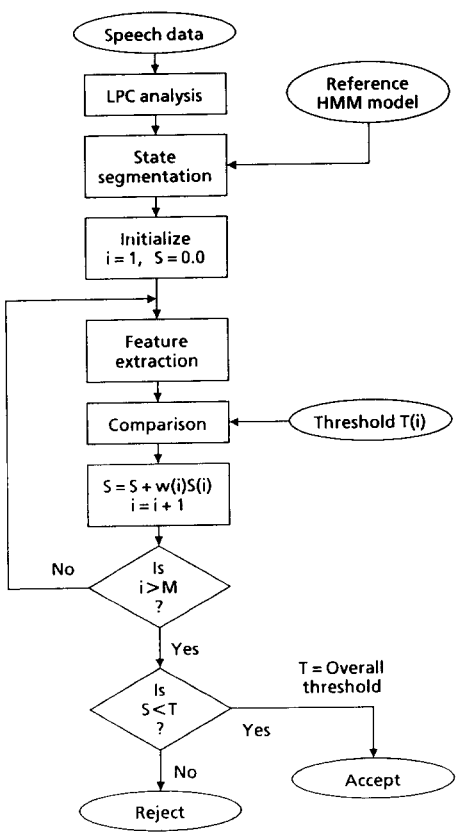

Fig. 7. Functional block diagram of the verification process in the HMM-based method. i: state index; S: distance score; T: threshold; w: weight; and M: total number of states.

The ASU-based speaker verification method described in Section 4.2 has been tested in the text-independent mode using the Naval Resource Management database recorded for DARPA (Rosenberg et al., 1990b). It has been shown that this approach can be extended to large vocabularies and continuous speech.

5.4. Neural net-based method

A new approach to speaker recognition, based on feed-forward neural models, has been investigated (Oglesby and Mason, 1990). Each person known to the system has a personalized neural net that is trained to be active only for that person's speech. It is assumed that including speech from many people in the training data of each net enables this approach to directly model differences between people's speech. It has been found that the chosen architecture and the amount of training strongly affect the recognition performance. Furthermore, the recognition performance has been shown to be comparable to that of the VQ approach based on personalized codebooks.

As an expansion of the VQ-based method, a connectionist approach has recently been developed based on the LVQ (learning vector quantization) algorithm (Bennani et al., 1990).

6. Speaker-adaptation techniques for speech recognition

6.1. Limits of speaker-independent methods

Recently, speaker-independent speech recognition methods using HMM techniques have been actively researched, and the recognition accuracy has been improved (Rabiner et al., 1983; Lee, 1988; Mariani, 1989). However, one of the disadvantages of the speaker-independent approach is that it neglects various useful characteristics of the speaker. 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 dis-

tributions of feature parameters are very broad or multi-modal, such as in the cases of the combination of male and female voices and of various dialects, it is difficult to separate phonemes using speaker-independent methods. To cope with these problems, it is essential to introduce speaker-adaptation techniques.

6.2. Classification of speaker-adaptation/normalization methods

Speaker adaptation or normalization ("speaker adaptation" indicates both adaptation and normalization hereafter) is the method of 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 short sentences. In large vocabulary recognition systems, training by the utterances of all vocabulary words is too troublesome for the users and consequently unrealistic. Therefore, training by a few words or short sentences is a practical and realistic solution.

Speaker-adaptation methods are generally classified into supervised (text-dependent) methods in which training words or sentences are known, and unsupervised methods in which arbitrary utterances can be used. Both methods can also be classified into off-line methods in which training words or sentences must be uttered before the recognition, and on-line 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 utterance of each speaker. Also, the system should actually adapt to the speaker's voice automatically using utterances for recognition. Such a system can be realized by the unsupervised, on-line adaptation mechanism. Humans have been found to have a mechanism of unsupervised, on-line speaker adaptation (Kato and Furui, 1985).

Since individual information (individuality) varies depending on the phoneme classes, it is effective to introduce the mechanism of explicitly or implicitly recognizing the phonemes included in the training utterances into the unsupervised adaptation process. By recognizing the phonemes and utilizing the phoneme-dependent individual information, adaptation performances can be improved. In such processes, it is crucially important that phoneme decision errors do not cause inappropriate adaptation.

6.3. Speaker cluster selection methods

One of the basic speaker-adaptation methods is the speaker cluster selection method. In this method, it is assumed that speakers can be divided into clusters, within which the speakers are similar. From many sets of phoneme template clusters representing speaker variability, the most suitable set for the new speaker is automatically selected.

In an HMM-based supervised method, a group of speakers is divided into speaker clusters, and a codebook and HMMs are generated for each cluster (Lee, 1988). Cluster-specific HMM parameters are derived by a training process using speakers in the cluster, or by a training process using all speakers and then converting the parameters to cluster-specific parameters through probabilistic mapping. Cluster-specific parameters for the cluster selected to be the most suitable for the speaker are used in the recognition stage as illustrated in Figure 8. It has been observed that when the size of training data is restricted, the possible number of clusters is limited, and therefore the adaptation performance is limited.

A neural network-based approach has also been investigated (Hampshire II and Waibel, 1990). In this approach, a multi-network Time-Delay Neural Network (TDNN)-based connectionist architecture performs multi-speaker phone discrimination (/b,d,g/). The overall network called "Meta–Pi network" gates the phonemic decisions of modules trained on individual speakers to form its overall classification decision as shown in Figure 9. In the Meta–Pi network, K speaker-dependent modules are linked by a multiplicative connection. The Meta–Pi paradigm implements a dynamically adaptive Bayesian MAP classifier, and it learns without supervision. By dynamically adapting to the input speech and focusing on a combination of speaker-specific modules, the net-

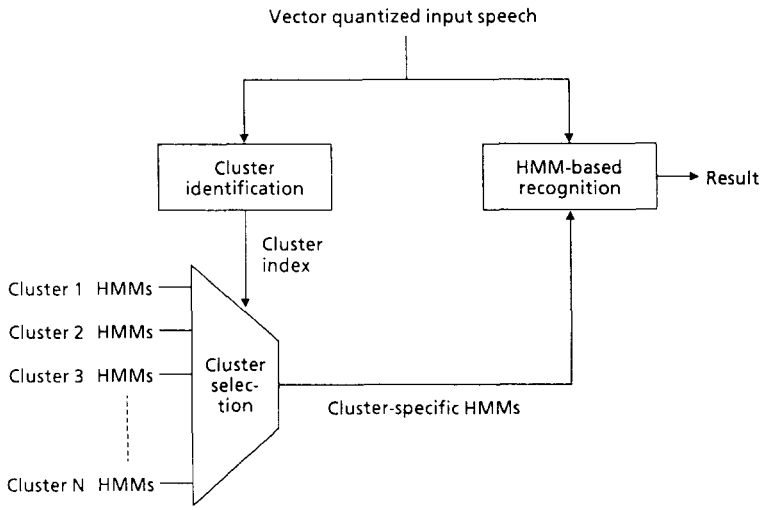

Fig. 8. Block diagram of speaker adaptation using the cluster selection method.

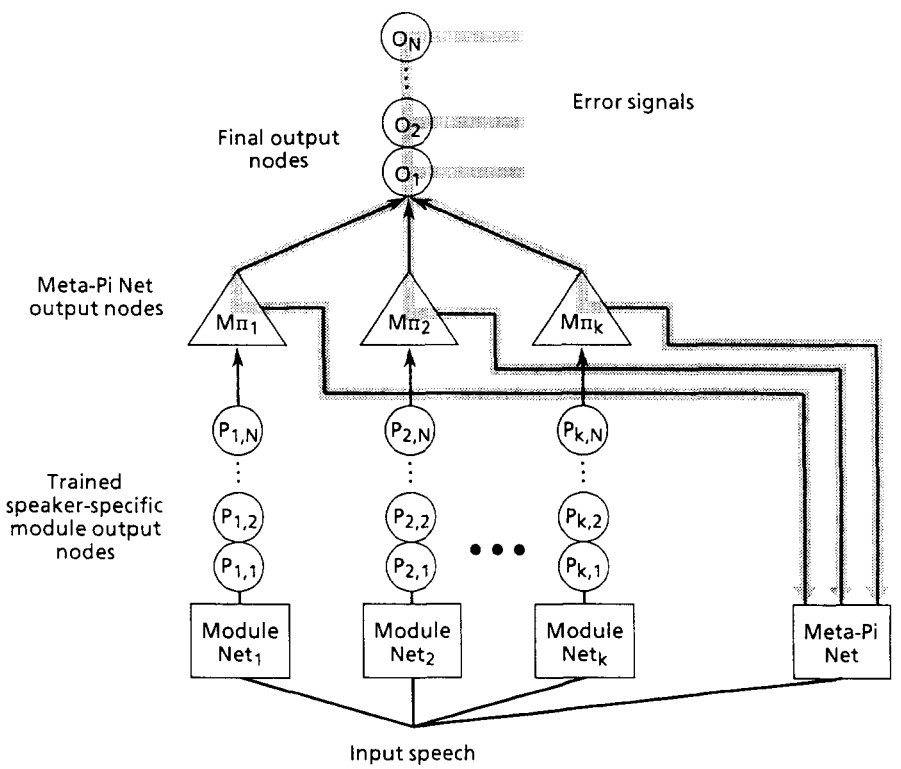

Fig. 9. The Meta–Pi network: K speaker-dependent modules linked by an integrating superstructure.

work outperforms a single TDNN trained on the speech of all K speakers.

6.4. Interpolated re-estimation algorithm

The deleted interpolation technique (Jelinek and Mercer, 1980) has recently been studied for generating speaker-adaptive parameters by interpolating speaker-independent, fixed existing parameters and speaker-dependent parameters created from a small number of training sentences. The speaker-independent parameters are well-trained but less appropriate for the individual speaker. On the other hand, the speaker-dependent parameters are appropriate for the individual speaker but poorly trained because of limited training data. The deleted interpolation technique can be used for combining the two parameter sets into estimates that are more suitable than speaker-independent parameters, yet more robust than speaker-dependent ones.

6.5. Codebook adaptation/normalization algorithm

Instead of interpolating speaker-independent and speaker-dependent parameters, speaker-adaptive parameters can be estimated from speaker-independent parameters based on mapping rules. The mapping rules are estimated from the relationship 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 (initial/reference) codebook to a new speaker or normalizing (adjusting) the input speech to the codebook have been proposed. Each word is represented in the word dictionary as single or multiple sequences of codebook entries. In the latter case, individual variations on how a word is uttered are modeled by multiple code sequences. The code sequences are not changed during the adaptation and are universally used for all speakers.

6.5.1. Supervised adaptation

In the case of supervised adaptation, the mapping rules are obtained through DTW or a forward–backward algorithm. Figure 10 is a block diagram of a supervised adaptation method (Shikano et al., 1986). The utterances by a reference speaker are initially used to create an initial codebook and these utterances are converted into sequences of codebook entries. In the training stage, training utterances by a new speaker are converted into code sequences and time-aligned with the same words or sentences uttered by the reference speaker, using the DTW technique. The spectral mapping function between the codebook elements of these two speakers is obtained based on the alignment functions, that is, the correspondence between the time axes. A histogram of correspondences between codebook elements of the reference speaker and the new speaker is calculated using the alignment results for all training words or sentences. The mapping

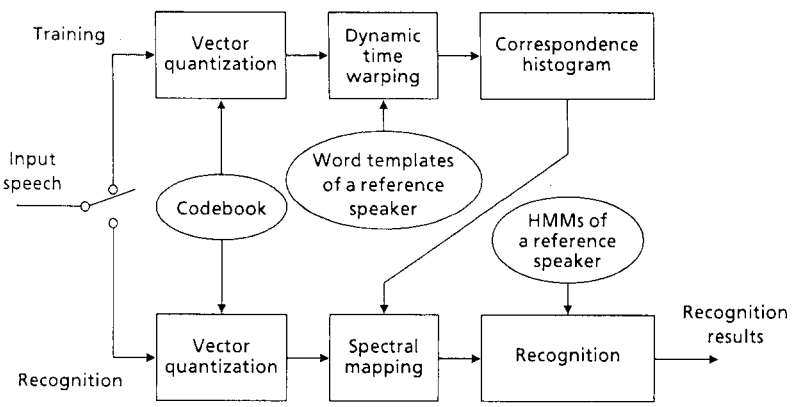

Fig. 10. Block diagram of supervised speaker adaptation by spectral mapping.

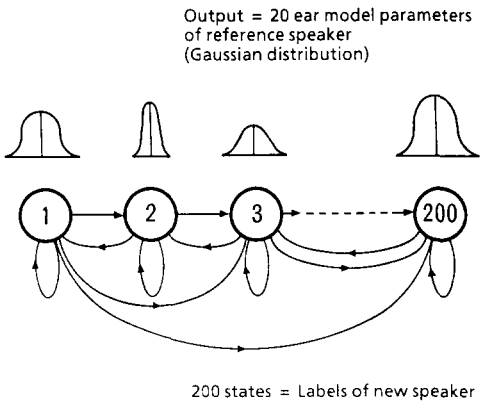

Fig. 11. Speaker Markov model describing the spectral differences of two speakers.

function is weighted by the histogram to find the best correspondence rule.

In the recognition stage, input speech is vector-quantized and mapped to the reference speaker's spectrum at every frame using the mapping rules. The similarity between mapped input speech and each word of the reference speaker is then calculated and used for the recognition decision. Experimental results show that recognition accuracies for 92-word and 422-word recognition are improved from 64 to 83% and from 48 to 81%, respectively, by this method.

In HMM-based recognition, HMM speech models derived from one "prototype" speaker are transformed so that they model the speech of the new speaker (Schwartz et al., 1987; Nishimura and Sugawara, 1988). This transformation is accomplished by probabilistic spectral mapping from one speaker's spectral space to that of another. The transformation matrix, which represents the conditional probability of a quantized spectrum of the new speaker, given the quantized spectrum of the reference speaker, is computed by applying a modified forward–backward algorithm to training utterances.

In the adaptive mode of the recognition system developed at BBN Laboratory, a 90.4% recognition rate was obtained using two-minute training utterances for the standard DARPA task with a grammar of a perplexity of 60 (Feng, 1989). The speaker-dependent performance with 28 minutes of training speech was 92.9%.

An approach to speaker adaptation for the IBM large-vocabulary HMM-based speech recognition system has also been investigated (Rigoll, 1989). The approach is based on the use of a stochastic model, called the "speaker Markov model", which is shown in Figure 11. The speaker Markov model indicates which prototype of the new speaker is likely to occur and what acoustic parameters are generated by the reference speaker if, at the same time, a certain prototype is generated by the new speaker. The average recognition rate dropped from 96.4% to only 95.2% for a 5000-word vocabulary task when the duration of training utterances was reduced from the usual 20 minutes to 5 minutes.

6.5.2. Supervised adaptation using neural networks

An adaptation method using neural networks, as shown in Figure 12, has also been studied (Iso

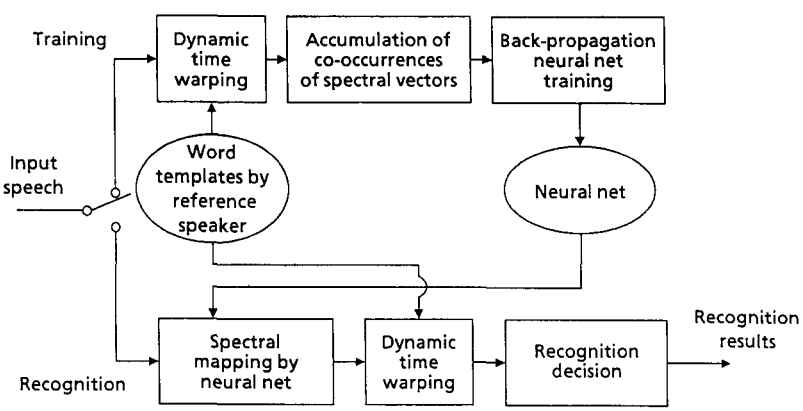

Fig. 12. Block diagram of speaker adaptation by neural network.

et al., 1989; Montacie et al., 1989). In this method, nonlinear continuous mapping between the reference spectral space and spectral space of a new speaker is represented by a multi-layer neural network.

The network training is executed using the back-propagation method so that the network generates reference template spectra at the output layer when corresponding training utterance spectra are given to the input layer. Corresponding data sets of reference templates and training utterance spectra are obtained using the DTW technique in the same way as the supervised adaptation method described in the previous subsection. VQ-distortion of training utterance spectra can be avoided and non-linear mapping can be realized by this neural network-based method. The problem, however, is that this method requires a large number of training words.

6.5.3. Unsupervised adaptation

Figure 13 is a block diagram of a word recognition system with an unsupervised codebook adaptation mechanism (Furui, 1989b, 1989c). The basic idea of this method is based on an adaptation algorithm for a segment vocoder (Shiraki and Honda, 1987). First, an initial codebook and a VQ-indexed word dictionary are prepared. The initial codebook is produced by clustering the voices of multiple speakers, and commonly serves as the initial condition for every new speaker.

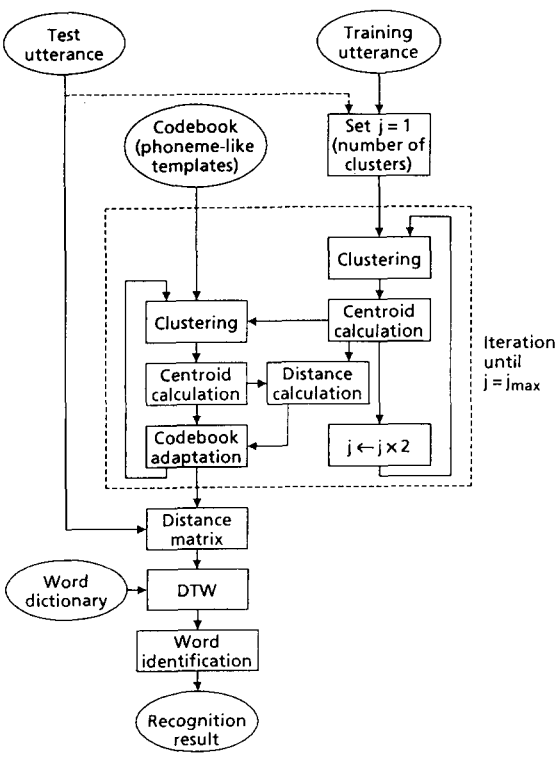

Fig. 13. Principal structure of a word recognition system with the unsupervised codebook adaptation mechanism.

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 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 as the weighted-sum of the deviation vectors of adjacent clusters. Figure 14 illustrates the hierarchical adaptation procedures of shifting the codebook elements from the beginning to the four-cluster stage. The size of the codebook is maintained throughout the adaptation process. Adaptation is thus performed hierarchically from global to local individuality.

This method was evaluated by a cepstrum-based 100-Japanese-word recognition system in which each word is represented by multiple se-

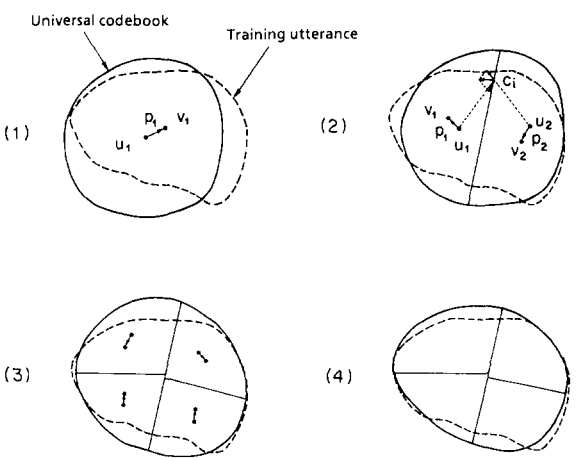

Fig. 14. Hierarchical codebook adaptation algorithm maintaining continuity between adjacent clusters. v_j: centroid of j-th training spectra cluster; u_j: centroid of corresponding codebook cluster; p_j: deviation vector; and c_i: shifting vector for i-th codebook element.

quences of codebook entries. Four speakers representing inter-speaker variability are selected and the voices of these speakers are used for creating an initial codebook and word dictionary. The codebook has 1024 entries. The multiple template method assures moderate recognition performance for new speakers at the beginning of the adaptation process. Results of experiments show that the adaptation using 10 arbitrary training words reduces the mean word recognition error rate from 4.9 to 2.9%. Since the error rate for speaker-dependent recognition is 2.2%, this method is highly effective. Several modifications of the adaptation method have also been investigated (Furui, 1989b).

6.6. *Reference template generation method*

A supervised adaptation procedure has been devised to generate phoneme templates adapted to each speaker using a set of transformation rules as well as the templates extracted from training words (Furui, 1980). Each new speaker is requested to provide utterances of only a fraction of the entire vocabulary as a training set. The transformation rules indicate the relationship between the templates for the entire vocabulary and those extracted from the training words. The rules are obtained through multiple regression analysis in a pretraining stage in which a group of speakers provides utterances of the entire vocabulary. The rules are universally used for all new speakers. The generated templates and speaker-independent templates are interpolated to create templates to be used in the recognition stage.

Results of recognition experiments using 67 Japanese airport names uttered by 30 male speakers have ascertained the effectiveness of this training procedure. A mean recognition accuracy of 98.2% was obtained after a 12-word training procedure.

7. Individuality problems in speech synthesis and coding

7.1. *Voice conversion*

Controlling or adding the individuality in synthesized speech is one of the important problems. Naturalness of synthesized voice is highly related to its individuality. This means that an ultimate goal of speech synthesis is the production of voices having the individuality of real human beings. Thus far no system has been constructed that can precisely control the synthesized voice quality or imitate a desired speaker's voice.

Along this line, a technique of converting voice quality from one speaker to another through vector quantization and spectrum mapping has been investigated (Abe et al., 1988). The basic idea of this technique is to make codebook mapping rules

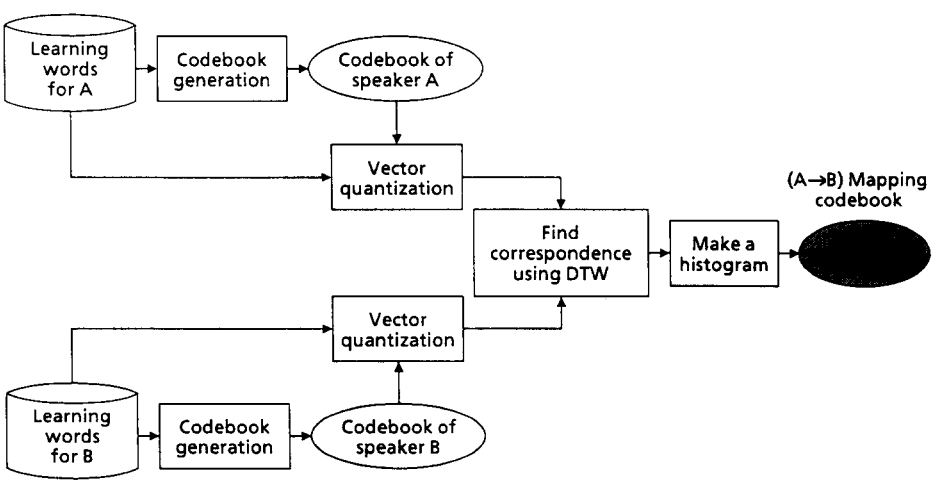

Fig. 15. Method for generating a mapping codebook for voice conversion.

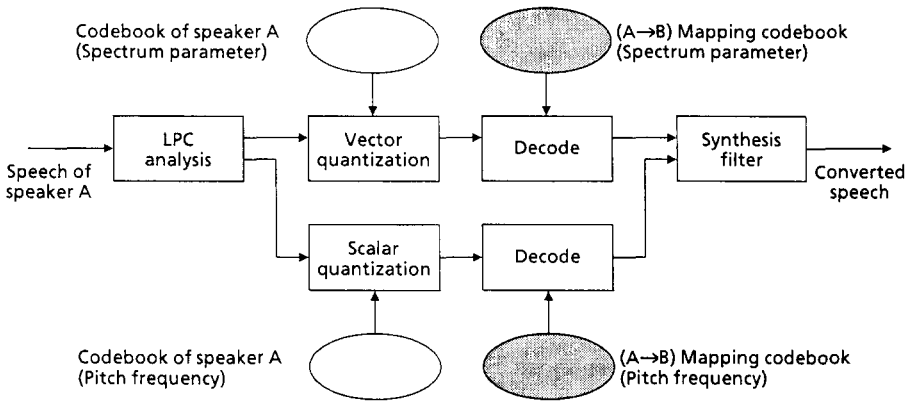

Fig. 16. Block diagram of voice conversion from speaker *A* to speaker *B*.

which represent the correspondence between different speakers' codebooks. A set of mapping rules for each codebook element is called a mapping codebook. Mapping codebooks for three parameter sets, that is, spectral parameters, energy values and pitch frequencies, are separately generated using training utterances.

The block diagram in Figure 15 illustrates the method of generating a mapping codebook for spectral parameters, which is similar to the technique used in the supervised speaker adaptation described in Section 6.5.1. First, a separate training word set is pronounced by each of two speakers, *A* and *B*, and vector-quantized frame by frame. The correspondence between vectors of the same words from the two speakers is determined using DTW. The vector correspondences between two speakers are accumulated throughout the training word set to create a histogram. Using the histogram as a weighting function, each element of the mapping codebook is defined as a weighted linear combination of the codebook elements.

Figure 16 shows a block diagram of a voice conversion from speaker *A* to speaker *B*. Synthesis is carried out by decoding (converting) the quantized parameters of the speaker *A* using the mapping codebooks between speakers *A* and *B*.

To evaluate the performance of this technique, hearing tests have been carried out under two kinds of voice conversion conditions. One is a conversion between male and female, the other is a conversion between male speakers. In the male-to-female conversion experiment, all converted utterances were judged as female, and in the male-to-male conversion, 65% of the converted voices were identified as the target speaker's voices.

7.2. Individuality problems in speech coding

In speech coding, the dependency of the coded speech quality on the individuality of the original speech increases with the more sophisticated, high-compression-rate methods. Robustness to the variation of voice individuality is one of the essential issues in developing and evaluating advanced speech coding methods. For this purpose, it is necessary to standardize speech database which cover a wide variety of voice individuality.

8. Conclusion

The algorithms for extracting individual information from speech waves, clarifying its mechanism, and controlling and normalizing the individuality are important future research topics for realizing advanced speech information processing systems, including speaker recognition, speech recognition, synthesis and coding systems. Among them, speaker recognition and speaker adaptation or normalization in speech recognition are particularly noteworthy research and development topics. Since these topics are two sides of the same problem: how best to separate the

speaker's information and the phonetic information in speech waves, algorithms for them should be investigated using a common approach.

For this purpose, it will be necessary to foster basic research on elucidating the mechanism of individuality in speech spectra and on inventing a sophisticated yet tractable model of speech variation based on a sufficiently large database.

References

M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1988), "Voice conversion through vector quantization", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, New York, S14.1.

Y. Bennani, F. Fogelman Soulie and P. Gallinari (1990), "A connectionist approach for automatic speaker identification", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, S5.2.

J. Eatock and J.S. Mason (1990), "Automatically focusing on good discriminating speech segments in speaker recognition", *Proc. Internat. Conf. Spoken Language Processing*, Kobe, 5.2.

M.W. Feng (1989), "Iterative normalization for speaker-adaptive training in continuous speech recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, S12.4.

S. Furui (1974), "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", *Trans. IECE*, 57-A, Vol. 12, pp. 880–887.

S. Furui (1980), "A training procedure for isolated word recognition systems", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, pp. 129–136.

S. Furui (1981), "Cepstral analysis technique for automatic speaker verification", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 29, No. 2, pp. 254–272.

S. Furui (1986), "Research on individuality features in speech waves and automatic speaker recognition techniques", *Speech Communication*, Vol. 5, No. 2, pp. 183–197.

S. Furui, (1989a), Digital Speech Processing, Synthesis, and Recognition (Marcel Dekker, New York).

S. Furui (1989b), "Unsupervised speaker adaptation method based on hierarchical spectral clustering", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, S6.9.

S. Furui (1989c), "Unsupervised speaker adaptation based on hierarchical spectral clustering", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-37, pp. 1923–1930.

S. Furui (1990), "Speaker-dependent feature extraction, speaker recognition, and speaker adaptation techniques", *Proc. VERBA 90*, pp. 164–170.

S. Furui, F. Itakura and S. Saito (1972), "Talker recognition by longtime averaged speech spectrum", *Trans. IECE*, 55-A, Vol. 10, pp. 549–556.

J.B. Hampshire II and A.H. Waibel (1990), "The Meta-Pi network: Connectionist rapid adaptation for high-performance multi-speaker phoneme recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, S3.9.

K. Iso, M. Asogawa, K. Yoshida and T. Watanabe (1989), "Speaker adaptation using neural network", *Proc. Spring Meeting of Acoust. Soc. Jap.*, 1-6-16 (in Japanese).

F. Jelinek and R.L. Mercer (1980), "Interpolated estimation of Markov source parameters from sparse data", in *Pattern Recognition in Practice*, ed. by E.S. Gelsema and L.N. Kanal (North-Holland, Amsterdam, the Netherlands), pp. 381–397.

B.-H. Juang and F.K. Soong (1990), "Speaker recognition based on source coding approaches", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, S5.4.

K. Kato and S. Furui (1985), "Listener adaptability for individual voice in speech perception", *Trans. Committee Hearing Res., Acoust. Soc. Japan*, H-85-5 (in Japanese).

K.-F. Lee (1988), Large-vocabulary speaker-independent continuous speech recognition: The SPHINX system, Ph.D. dissertation, Computer Science Department, Carnegie Mellon University.

K.P. Li and E.H. Wrench, Jr. (1983), "An approach to text-independent speaker recognition with short utterances", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Boston, 12.9.

J. Mariani (1989), "Recent advances in speech processing", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, S9.1.

T. Matsui and S. Furui (1990), "Text-independent speaker recognition using vocal tract and pitch information", *Proc. Internat. Conf. Spoken Language Processing*, Kobe, 5.3.

T. Matsui and S. Furui (1991), "A text-independent speaker recognition method robust against utterance variations", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Toronto, Canada.

C. Montacie, K. Choukri and G. Chollet (1989), "Speech recognition using temporal decomposition and multi-layer feed-forward automata", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, S8.6.

J.M. Naik, L.P. Netsch and G.R. Doddington (1989), "Speaker verification over long distance telephone lines", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, S10b.3.

M. Nishimura and K. Sugawara (1988), "Speaker adaptation method for HMM-based speech recognition", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, New York, S5.7.

J. Oglesby and J.S. Mason (1990), "Optimization of neural models for speaker identification", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, S5.1.

A.B. Poritz (1982), "Linear predictive hidden Markov models and the speech signal", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Paris, S11.5.

L.R. Rabiner, S.E. Levinson and M.M. Sondhi (1983), "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition", *Bell Syst. Techn. J.*, Vol. 62, No. 4, pp. 1075–1105.

G. Rigoll (1989), "Speaker adaptation for large vocabulary speech recognition systems using 'speaker Markov models' ", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Glasgow, Scotland, S1.2.

A.E. Rosenberg, C.-H. Lee and F.K. Soong (1990a), "Subword unit talker verification using hidden Markov models", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, S5.3.

A.E. Rosenberg, C.-H. Lee, F.K. Soong and M.A. McGee (1990b), "Experiments in automatic talker verification using sub-word unit hidden Markov models", *Proc. Internat. Conf. Spoken Language Processing*, Kobe, 5.4.

M. Savic and S.K. Gupta (1990), "Variable parameter speaker verification system based on hidden Markov modeling", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Albuquerque, S5.7.

R. Schwartz, Y.-L. Chow and F. Kubala (1987), "Rapid speaker adaptation using a probabilistic spectral mapping", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Dallas, 15.3.

K. Shikano, K.-F. Lee and R. Reddy (1986), "Speaker adaptation through vector quantization", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, Tokyo, 49.5.

Y. Shiraki and M. Honda (1987), "Speaker adaptation algorithms for segment vocoder", *Trans. Committee of Speech Res., Acoust. Soc. Japan*, SP87-67 (in Japanese).

F.K. Soong and A.E. Rosenberg (1988), "On the use of instantaneous and transitional spectral information in speaker recognition", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 36, No. 6, pp. 871–879.

Y.-C. Zheng and B.-Z. Yuan (1988), "Text-dependent speaker identification using circular hidden Markov models", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, New York, S13.3.