Encyclopedia of Language and Linguistics, 2nd Ed., Edited by Keith Brown
Elsevier: Oxford, pp. 619-629, 2006, ISBN 0-08-044299-4

Automatic Speaker Recognition and Verification

Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology
2-12-1, O-okayama, Meguro-ku, Tokyo, 152 Japan
furui@cs.titech.ac.jp

Abstract

Speaker recognition is the process of recognizing the speaker using speech signals, which can be classified into speaker identification and verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Speaker recognition can also be classified into text-dependent, text-independent and text-prompted methods. Spectral envelope and prosody features of speech are usually used as speaker features. In order to accommodate intra-speaker variations of signal characteristics, it is important to apply parameter-domain and/or likelihood-domain normalization/adaptation techniques. High-level features, such as word idiolect, pronunciation, phone usage and prosody, have recently been investigated in text-independent speaker verification.

Keywords: speaker identification, speaker verification, text-dependent, text-independent, text-prompted, normalization, adaptation, high-level features, cepstral coefficients, prosody

1. Principles of Speaker Recognition

1.1 General Principles and Applications

Speaker recognition is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves (Furui, 1989; Furui, 1997; Rosenberg and Soong, 1991) be used to verify identities being claimed by people accessing systems; that is, it enables access control of various services by voice. Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access of computers. Another important application of speaker recognition technology is use for forensic purposes (Kunzel, 1994).

Speaker identity is correlated with physiological and behavioral characteristics of the speech production system of an individual speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech. The most commonly used short-term spectral measurements are FFT/LPC-derived cepstral coefficients and their regression coefficients (Furui, 1981). As for the regression coefficients, typically, the first- and second-order coefficients, that is, derivatives of the time functions of cepstral coefficients, are extracted at every frame period to represent spectral dynamics. They are respectively called the delta-cepstral and delta-delta-cepstral coefficients.

1.2 Speaker Identification and Verification

Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification is the process of determining from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the identity claimed by a speaker. Most of the applications in which voice is used to confirm the identity of a speaker are classified as speaker verification.

The basic elements of a speaker recognition system are shown in Fig. 1. In the speaker identification task, a speech utterance from an unknown speaker is analyzed and compared with speech models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. In speaker verification, an identity is claimed by an unknown speaker, and an utterance of this unknown speaker is compared with a model for the speaker whose identity is being claimed. If the match is good enough, that is, above a threshold, the identity claim is accepted. A high threshold makes it difficult for impostors to be accepted by the system, but with the risk of falsely rejecting valid users. Conversely, a low threshold enables valid users to be accepted consistently, but with the risk of accepting impostors. To set the threshold at the desired level of customer rejection (false rejection) and impostor acceptance (false acceptance), data showing distributions of customer and impostor scores are necessary.

Fig. 1 – Basic structures of speaker recognition systems.

The fundamental difference between identification and verification is the number of decision alternatives. In identification, the number of decision alternatives is equal to the size of the population, whereas in verification there are only two choices, acceptance or rejection, regardless of the population size. Therefore, speaker identification performance decreases as the size of the population increases, whereas speaker verification performance approaches a constant independent of the size of the population, unless the distribution of physical characteristics of speakers is extremely biased.

There is also a case called "open set" identification, in which a reference model for an unknown speaker may not exist. In this case, an additional decision alternative, "the unknown does not match any of the models", is required. Verification can be considered a special case of the "open set" identification mode in which the known population size is one. In either verification or identification, an additional threshold test can be applied to determine whether the match is sufficiently close to accept the decision, or if not, to ask for a new trial.

The effectiveness of speaker verification systems can be evaluated by using receiver operating characteristics (ROC) curve adopted from psychophysics. The ROC curve is obtained by assigning two probabilities, the probability of correct acceptance (1 – false rejection rate) and the probability of incorrect acceptance (false acceptance rate), to the vertical and horizontal axes respectively, and varying the decision threshold (Furui, 1989). The detection error trade-off (DET) curve has recently become popular, in which false rejection and false acceptance rates are assigned to the vertical and horizontal axes respectively (Bonastre et al., 2004) . The error curve is usually plotted on a normal deviate scale. With this scale, a speaker recognition system whose true speaker and impostor scores are Gaussians with the same variance will result in a linear curve with a slope equal to – 1. The DET curve representation is therefore more easily readable than the ROC curve and allows for a comparison of the system's performances over a large range of operating conditions.

The equal-error rate (EER) is a commonly accepted overall measure of system performance. It corresponds to the threshold at which the false acceptance rate is equal to the false rejection rate.

1.3 Text-Dependent, Text-Independent and Text-Prompted Methods

Speaker recognition methods can also be divided into text-dependent (fixed passwords) and text-independent (no specified passwords) methods. The former require the speaker to provide utterances of key words or sentences, the same text being used for both training and recognition, whereas the latter do not rely on a specific text being spoken. The text-dependent methods are usually based on template/model-sequence-matching techniques in which the time axes of an input speech sample and reference templates or reference models of the registered speakers are aligned, and the similarities between them are accumulated from the beginning to the end of the utterance. Since this method can directly exploit voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent method.

There are several applications, such as forensic (Kunzel, 1994) and surveillance applications, in which predetermined key words cannot be used. Moreover, human beings can recognize speakers irrespective of the content of the utterance. Therefore, text-independent methods have recently attracted more attention. Another advantage of text-independent recognition is that it can be done sequentially, until a desired significance level is reached, without the annoyance of the speaker having to repeat key words again and again.

Both text-dependent and independent methods have a serious weakness. That is, these security systems can easily be circumvented, because someone can play back the recorded voice of a registered speaker uttering key words or sentences into the microphone and be accepted as the registered speaker. Another problem is that people often do not like text-dependent systems because they do not like to utter their identification number, such as their social security number, within the hearing of other people. To cope with these problems, some methods use a small set of words, such as digits as key words, and each user is prompted to utter a given sequence of key words which is randomly chosen every time the system is used (Higgins et al., 1991; Rosenberg et al., 1987). Yet even this method is not reliable enough, since it can be circunvented with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted speaker recognition method has been proposed in which password sentences are completely changed every time (See Section 3).

2 Text-Dependent Speaker Recognition Methods

Text-dependent speaker recognition methods can be classified into DTW (dynamic time warping) or HMM (hidden Markov model) based methods.

2.1 DTW-Based Methods

In this approach, each utterance is represented by a sequence of feature vectors, generally, short-term spectral feature vectors, and the trial-to-trial timing variation of utterances of the same text is normalized by aligning the analyzed feature vector sequence of a test utterance to the template feature vector sequence using a DTW algorithm. The overall distance between the test utterance and the template is used for the recognition decision. When multiple templates are used, distances between the test utterance and the templates are averaged and used for the decision.

2.2 HMM-Based Methods

An HMM can efficiently model the statistical variation in spectral features. Therefore, HMM-based methods have achieved significantly better recognition accuracies than DTW-based methods (Naik et al., 1989).

3 Text-Independent Speaker Recognition Methods

In text-independent speaker recognition, generally the words or sentences used in recognition trials cannot be predicted. Since it is impossible to model or match speech events at the word or sentence level, the following three kinds of methods shown in Fig. 2 are being actively investigated (Furui, 1997).

Fig. 2 - Basic structures of text-independent speaker recognition methods.

3.1 Long-Term-Statistics-Based Methods

As text-independent features, long-term sample statistics of various spectral features, such as the mean and variance of spectral features over a series of utterances, have been used (Fig. 2(a)). However, long-term spectral averages are extreme condensations of the spectral characteristics of a speaker's utterances and, as such, lack the discriminating power of the sequences of short-term spectral features used as models in text-dependent methods. In one of the trials using the long-term averaged spectrum (Furui et al., 1972), the effect of session-to-session variability was reduced by introducing a weighted cepstral distance measure.

Studies on using statistical dynamic features have also been carried out. Multivariate auto-regression (MAR) models have been applied to the time series of cepstral vectors to characterize speakers, and good speaker recognition results have been obtained (Montacie et al., 1992; Griffin et al., 1994). It was reported that the optimum order of the MAR model was 2 or 3, and that distance normalization using *a posteriori* probability was essential to obtain good results in speaker verification.

3.2 VQ-Based Methods

A set of short-term training feature vectors of a speaker can be used directly to represent the essential characteristics of that speaker. However, such a direct representation is impractical when the number of training vectors is large, since the memory and amount of computation required become prohibitively large. Therefore, attempts have been made to find efficient ways of compressing the training data using vector quantization (VQ) techniques.

In this method, VQ codebooks, consisting of a small number of representative feature vectors, are used as an efficient means of characterizing speaker-specific features (Li and Wrench, 1983; Matsui and Furui, 1990; Matsui and Furui, 1991; Rosenberg and Soong, 1987). In the recognition stage, an input utterance is vector-quantized by using the codebook of each reference speaker; the VQ distortion accumulated over the entire input utterance is used for making the recognition determination (Fig. 2(b)).

Matsui et al. (Matsui and Furui, 1990; Matsui and Furui, 1991) tried a method using a VQ-codebook for long feature vectors consisting of instantaneous and transitional features calculated for both cepstral coefficients and fundamental frequency. Since the fundamental frequency cannot be extracted from unvoiced speech, there are two separate codebooks for voiced and unvoiced speech for each speaker. A new distance measure was introduced to take into account intra- and inter-speaker variability and to deal with the problem of outlier in the distribution of feature vectors. The outlier vectors correspond to intersession spectral variation and to the difference between phonetic content of the training texts and the test utterances. It was confirmed that, although the fundamental frequency achieved only a low recognition rate by itself, the recognition accuracy was greatly improved by combining the fundamental frequency with spectral envelope features.

In contrast with the memoryless VQ-based method, non-memoryless source coding algorithms have also been studied using a segment (matrix) quantization technique (Juang and Soong, 1990). The advantage of a segment quantization codebook over a VQ codebook representation is its characterization of the sequential nature of speech events. Higgins and Wohlford (Higgins and Wohlford, 1986) proposed a segment modeling procedure for constructing a set of representative time normalized segments, which they called "filler templates". The procedure, a combination of K-means clustering and dynamic programming time alignment, provided a means for handling temporal variation.

3.3 Ergodic-HMM-Based Methods

The basic structure is the same as the VQ-based method, but in this method an ergodic HMM is used instead of a VQ codebook. Over a long timescale, the temporal variation in speech signal parameters is represented by stochastic Markovian transitions between states. Poritz (Poritz, 1982) proposed using a five-state ergodic HMM (i.e., all possible transitions between states are allowed) to classify speech segments into one of the broad phonetic categories corresponding to the HMM states. A linear predictive HMM was adopted to characterize the output probability function. He characterized the automatically obtained categories as strong voicing, silence, nasal/liquid, stop burst/post silence, and frication.

Gauvain et al. (Gauvain et al., 1995) investigated a statistical modeling approach, where each speaker was viewed as a source of phonemes, modeled by a fully connected Markov chain. Maximum *a posteriori* (MAP) estimation was used to generate speaker-specific models from a set of speaker-independent seed models. The lexical and syntactic structures of the language were approximated by local phonotactic constraints. The unknown speech is recognized by all of the speakers' models in parallel, and the hypothesized identity is that associated with the model set having the highest likelihood. Since phonemes and speakers are simultaneously recognized by using speaker-specific Markov chains, this method can be considered as an extension of the ergodic-HMM-based method.

Matsui et al. (Matsui and Furui, 1992) compared the VQ-based method with the discrete/continuous ergodic HMM-based method, particularly from the viewpoint of robustness against utterance variations. They found that the continuous ergodic HMM method is far superior to the discrete ergodic HMM method and that the continuous ergodic HMM method is as robust as the VQ-based method when enough training data is available. However, when little data is available, the VQ-based method is robuster than the continuous HMM method. They investigated speaker identification rates using the continuous HMM as a function of the number of states and mixtures. It was shown that the speaker recognition rates were strongly correlated with the total number of mixtures, irrespective of the number of states. This means that using information on transitions between different states is ineffective for text-independent speaker recognition.

Rose et al. (Rose and Reynolds, 1990) investigated a technique based on maximum likelihood estimation of a Gaussian mixture model (GMM) representation of speaker identity. This method corresponds to the single-state continuous ergodic HMM. Gaussian mixtures are noted for their robustness as a parametric model and for their ability to form smooth estimates of rather arbitrary underlying densities. The VQ-based method can be regarded as a special (degenerate) case of a single-state HMM with a distortion measure being used as the observation probability.

3.4 Speech-Recognition-Based Methods

The VQ- and HMM-based methods can be regarded as methods that use phoneme-class-dependent speaker characteristics contained in short-term spectral features through implicit phoneme-class recognition. In other words, phoneme-classes and speakers are simultaneously recognized in these methods. On the other hand, in the speech-recognition-based methods (Fig. 2(c)), phonemes or phoneme-classes are explicitly recognized, and then each phoneme (-class) segment in the input speech is compared with speaker models or templates corresponding to that phoneme (-class).

Savic et al. (Savic and Gupta, 1990) used a five-state ergodic linear predictive HMM for broad phonetic categorization. In their method, after frames that belong to particular phonetic categories have been identified, feature selection is performed. In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after phonetic categorization, a comparison with the reference template for each particular category provides a verification score for that category. The final verification score is a weighted linear combination of the scores for each category. The weights are chosen to reflect the effectiveness of particular categories of phonemes in discriminating between speakers and are adjusted to maximize the verification performance. Experimental results showed that verification accuracy can be considerably improved by this category-dependent weighted linear combination method.

Rosenberg et al. have been testing a speaker verification system using 4-digit phrases in actual field conditions with a banking application (Rosenberg et al., 1991) input speech is segmented into individual digits using a speaker-independent HMM. The frames within the word boundaries for a digit are compared with the corresponding speaker-specific HMM digit model and the Viterbi likelihood score is computed. This is done for each of the digits making up the input utterance. The verification score is defined to be the average normalized log-likelihood score over all the digits in the utterance.

Newman et al. (Newman et al., 1996) used a large vocabulary speech recognition system for speaker verification. A set of speaker-independent phoneme models were adapted to each speaker. Speaker verification consisted of two stages. First, speaker-independent speech recognition was run on each of the test utterances to obtain phoneme segmentation. In the second stage, the segments were scored against the adapted models for a particular target speaker. The scores were normalized by those with speaker-independent models. The system was evaluated using the 1995 NIST-administered speaker verification database, which consists of data taken from the Switchboard corpus. The results showed that this method could not out-perform Gaussian mixture models.

4 Text-Prompted Speaker Recognition

In this method, key sentences are completely changed every time (Matsui and Furui, 1993; Matsui and Furui, 1994b). The system accepts the input utterance only when it determines that the registered speaker uttered the prompted sentence. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. This method not only accurately recognizes speakers, but can also reject an utterance whose text differs from the prompted text, even if it is uttered by a registered speaker. Thus, a recorded and played back voice can be correctly rejected.

This method uses speaker-specific phoneme models as basic acoustic units. One of the major issues in this method is how to properly create these speaker-specific phoneme models when using training utterances of a limited size. The phoneme models are represented by Gaussian-mixture continuous HMMs or tied-mixture HMMs, and they are made by adapting speaker-independent phoneme models to each speaker's voice.

In the recognition stage, the system concatenates the phoneme models of each registered speaker to create a sentence HMM, according to the prompted text. Then the likelihood of input speech against the sentence model is calculated and used for the speaker verification determination.

5 Normalization and Adaptation Techniques

How can we normalize intra-speaker variation of likelihood (similarity) values in speaker verification? The most significant factor affecting automatic speaker recognition performance is variation in signal characteristics from trial to trial (intersession variability or variability over time). Variations arise from the speaker him/herself, from differences in recording and transmission conditions, and from noise. Speakers cannot repeat an utterance precisely the same way from trial to trial. It is well known that samples of the same utterance recorded in one session are much more highly correlated than tokens recorded in separate sessions. There are also long term trends in voices (Furui et al, 1972; Furui, 1974).

It is important for speaker recognition systems to accommodate these variations. Adaptation of the reference model as well as the verification threshold for each speaker is indispensable to maintain a high recognition accuracy for a long period. In order to compensate for the variations, two types of normalization techniques have been tried — one in the parameter domain, and the other in the distance/similarity domain. The latter technique uses the likelihood ratio or *a posteriori* probability. To adapt HMMs for noisy conditions, various techniques including the HMM composition (PMC: parallel model combination) method, have proved successful.

5.1 Parameter-Domain Normalization

As one typical normalization technique in the parameter domain, spectral equalization, the so-called "blind equalization" method, has been confirmed to be effective in reducing linear channel effects and long-term spectral variation (Atal et al., 1974; Furui, 1981). This method is especially effective for text-dependent speaker recognition applications using sufficiently long utterances. In this method, cepstral coefficients are averaged over the duration of an entire utterance, and the averaged values are subtracted from the cepstral coefficients of each frame (CMS; cepstral mean subtraction). This method can compensate fairly well for additive variation in the log spectral domain. However, it unavoidably removes some text-dependent and speaker-specific features, so it is inappropriate for short utterances in speaker recognition applications. It was shown that time derivatives of cepstral coefficients (delta-cepstral coefficients) are resistant to linear channel mismatch between training and testing (Furui, 1981; Soong and Rosenberg, 1988).

5.2 Likelihood Normalization

Higgins et al. (Higgins et al., 1991) proposed a normalization method for distance (similarity or likelihood) values that uses a likelihood ratio. The likelihood ratio is the ratio of the conditional probability of the observed measurements of the utterance given the claimed identity is correct to the conditional probability of the observed measurements given the speaker is an impostor (normalization term). Generally, a log-likelihood ratio indicates a valid claim, whereas a negative value indicates an imposter. The likelihood ratio normalization approximates optimal scoring in Bayes' sense.

This normalization method is, however, unrealistic because conditional probabilities must be calculated for all the reference speakers, which requires large computational cost. Therefore, a set of speakers, "cohort speakers", who are representative of the population distribution near the claimed speaker has been chosen for calculating the normalization term. Another way of choosing the cohort speaker set is to use speakers who are typical of the general population. Reynolds (Reynolds, 1994) reported that a randomly selected, gender-balanced background speaker population outperformed a population near the claimed speaker.

Matsui et al. (Matsui and Furui, 1993; Matsui and Furui, 1994a) proposed a normalization method based on *a posteriori* probability. The difference between the normalization method based on the likelihood ratio and that based on *a posteriori* probability is whether or not the claimed speaker is included in the impostor speaker set for normalization; the cohort speaker set in the likelihood-ratio-based method does not include the claimed speaker, whereas the normalization term for the *a posteriori*-probability-based method is calculated by using a set of speakers including the claimed speaker. Experimental results indicate that both normalization methods almost equally improve speaker separability and reduce the need for speaker-dependent or text-dependent thresholding, compared with scoring using only the model of the claimed speaker (Matsui and Furui, 1994a; Rosenberg, 1992).

Carey et al. (Carey and Parris, 1992) proposed a method in which the normalization term is approximated by the likelihood for a world model representing the population in general. This method has an advantage that the computational cost for calculating the normalization term is much smaller than the original method since it does not need to sum the likelihood values for cohort speakers. Matsui et al. (Matsui and Furui, 1994a) proposed a method based on tied-mixture HMMs in which the world model is made as a pooled mixture model representing the parameter distribution for all the registered speakers.

The use of a single background model for calculating the normalization term has become the predominate approach used in speaker verification systems.

Since these normalization methods neglect absolute deviation between the claimed speaker's model and the input speech, they cannot differentiate highly dissimilar speakers. Higgins et al. (Higgins et al.,1991) reported that a multilayer network decision algorithm makes effective use of the relative and absolute scores obtained from the matching algorithm.

A family of normalization techniques has recently been proposed, in which the scores are normalized by subtracting the mean and then dividing by standard deviation, both terms having been estimated from the (pseudo) imposter score distribution. Different possibilities are available for computing the imposter score distribution: Znorm, Hnorm, Tnorm, Htnorm, Cnorm and Dnorm (Bimbot et al., 2004). The state-of-the-art text-independent speaker verification techniques associate one or several parameterization level normalization (CMS, feature variance normalization, feature warping, etc.) with a world model normalization and one or several score normalizations.

5.3 HMM Adaptation for Noisy Conditions

Increasing the robustness of speaker recognition techniques against noisy speech or speech distorted by a telephone is a crucial issue in real applications. Rose et al. (Rose et al., 1994) applied the HMM composition (PMC) method (Gales and Young, 1993; Martin et al., 1993) to speaker identification under noisy conditions. The HMM composition is a technique to combine a clean speech HMM and a background noise HMM to create a noise-added speech HMM. In order to cope with the problem of the variation of the signal-to-noise ratio (SNR), Matsui et al. (Matsui and Furui, 1996b) proposed a method in which several noise-added HMMs with various SNRs were created and the HMM that had the highest likelihood value for the input speech was selected. A speaker decision was made using the likelihood value corresponding to the selected model. Experimental application of this method to text-independent speaker identification and verification in various kinds of noisy environments demonstrated considerable improvement in speaker recognition.

5.4 Updating Models and *A Priori* Threshold for Speaker Verification

How to update speaker models to cope with the gradual changes in people's voices is an important issue. Since we cannot ask every user to utter many utterances at many different sessions in real situations, it is necessary to build each speaker model based on a small amount of data collected in a few sessions, and then the model must be updated using speech data collected when the system is used.

How to set the *a priori* decision threshold for speaker verification is another important issue. In most laboratory speaker recognition experiments, the threshold is set *a posteriori* so that the equal error rate (EER) is achieved. Since the threshold cannot be set *a posteriori* in real situations, we have to have practical ways to set the threshold before verification. It must be set according to the relative importance of the two errors, which depends on the application.

These two problems are intrinsically related each other. Furui (Furui, 1981) proposed methods for updating reference templates and the threshold in DTW-based speaker verification. An optimum threshold was estimated based on the distribution of overall distances between each speaker's reference template and a set of utterances of other speakers (interspeaker distances). The interspeaker distance distribution was approximated by a normal distribution, and the threshold was calculated by the linear combination of its mean value and standard deviation. The intraspeaker distance distribution was not taken into account in the calculation, mainly because it is difficult to obtain stable estimates of the intraspeaker distance distribution from small numbers of training utterances. The reference template for each speaker was updated by averaging new utterances and the present template after time registration. Matsui et al. (Matsui and Furui, 1996a) extended these methods and applied them to text-independent and ext-prompted speaker verification using HMMs.

6. High-level Speaker Recognition

Recently, high-level features such as word idiolect, pronunciation, phone usage, prosody, etc. have been successfully used in text-independent speaker verification. Typically, high-level-feature recognition systems produce a sequence of symbols from the acoustic signal and then perform recognition using the frequency and co-occurrence of symbols. In Doddington's idiolect work (Doddington, 2001), word unigrams and bigrams from manually transcribed conversations were used to characterize a particular speaker in a traditional target/background likelihood ratio framework. Campbell et al. (Campbell et al., 2004) proposed the use of support vector machines for performing the speaker verification task based on phone and word sequences obtained using phone recognizers. The benefit of these features was demonstrated in the "NIST extended data" task for speaker verification; with enough conversational data, a recognition system can become "familiar" with a speaker and achieve excellent accuracy. The corpus was a combination of phases 2 and 3 of the Switchboard-2 corpora. Each training utterance in the corpus consisted of a conversation side that was nominally of length 5 minutes (approximately 2.5 minutes of speech) recorded over a land-line telephone. Speaker models were trained using 1 – 16 conversation sides. These methods need utterances of at least several minutes long, much longer than those used in conventional speaker recognition methods.

7. Relation with Other Speech Technology

Speaker characterization techniques are related to research on improving speech recognition accuracy by speaker adaptation (Furui, 1991), improving synthesized speech quality by adding the natural characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another. Studies on automatically extracting the speech periods of each person separately from a dialogue/conversation/meeting involving more than two people have appeared as an extension of speaker recognition technology (Gish et al., 1991; Sin et al., 1992; Wilcox et al., 1994). Increasingly, speaker segmentation and clustering techniques are being used to aid adapting speech recognizers and for supplying metadata for audio indexing and searching.

References

Atal, B. S. (1974) "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," J. Acoust. Soc. Am., Vol. 55, No. 6, pp. 1304-1312.

Bimbot, F. J., Bonastre, F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz D. and Reynolds, D. A. (2004) "A Tutorial on Text-Independent Speaker Verification," EURASIP Journ. on Applied Signal Processing, pp. 430-451.

Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A. and Leek, T. R. (2004) "High-Level Speaker Verification with Support Vector Machines," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. I-73-76.

Carey, M. J. and Parris, E. S. (1992) "Speaker Verification Using Connected Words," Proc. Institute of Acoustics, Vol. 14, Part 6, pp. 95-100.

Doddington, G. (2001) "Speaker Recognition Based on Idiolectal Differences between Speakers," Proc. Eurospeech, pp. 2521-2524.

Furui, S., Itakura, F., and Saito, S. (1972) "Talker Recognition by Longtime Averaged Speech Spectrum," Trans. IECE, 55-A, Vol. 1, No. 10, pp. 549-556.

Furui, S. (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.

Furui, S. (1981) "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoust. Speech, Signal Processing, Vol. 29, No. 2, pp. 254-272.

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

Furui, S. (1991) "Speaker-Independent and Speaker-Adaptive Recognition Techniques," in Furui, S. and Sondhi, M. M. (Eds.) *Advances in Speech Signal Processing*, New York: Marcel Dekker, pp. 597-622.

Furui, S. (1997) "Recent Advances in Speaker Recognition", Proc. First Int. Conf. Audio- and Video-based Biometric Person Authentication, Crans-Montana, Switzerland, pp. 237-252.

Gales, M. J. F. and Young, S. J. (1993) "HMM Recognition in Noise Using Parallel Model Combination," Proc. Eurospeech, Berlin, pp. II-837-840.

Gauvain, J. L., Lamel, L. F. and Prouts, B. (1995) "Experiments with Speaker Verification over the Telephone," Proc. Eurospeech, Madrid, pp. 651-654.

Griffin, C. T. Matsui and Furui, S. (1994) "Distance Measures for Text-Independent Speaker Recognition Based on MAR Model," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Adelaide, 23.6, pp. I-309-312.

Gish, H., Siu, M. and Rohlicek, R. (1991) "Segregation of Speakers for Speech Recognition and Speaker

Identification," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S13.11, pp. 873-876.

Higgins, A. L. and Wohlford, R. E. (1986) "A New Method of Text-Independent Speaker Recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 17.3, pp. 869-872.

Higgins, A., Bahler L. and Porter, J. (1991) "Speaker Verification Using Randomized Phrase Prompting," Digital Signal Processing, Vol. 1, pp. 89-106.

Juang, B. -H. and Soong, F. K. (1990) "Speaker Recognition Based on Source Coding Approaches," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.4, pp. 613-616.

Kunzel, H. J. (1994) "Current Approaches to Forensic Speaker Recognition," Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp.135-141.

Li, K. -P. and Wrench Jr., E. H. (1983) "An Approach to Text-Independent Speaker Recognition with Short Utterances," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 12.9, pp. 555-558.

Martin, F., Shikano, K. and Minami, Y. (1993) "Recognition of Noisy Speech by Composition of Hidden Markov Models," Proc. Eurospeech, Berlin, pp. II-1031-1034.

Matsui T. and Furui, S. (1990) "Text-Independent Speaker Recognition Using Vocal Tract and Pitch Information," Proc. Int. Conf. Spoken Language Processing, Kobe, 5.3, pp. 137-140.

Matsui T. and Furui, S. (1991) "A Text-Independent Speaker Recognition Method Robust Against Utterance Variations," Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, S6.3, pp. 377-380.

Matsui T. and Furui, S. (1992) "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMMs," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, San Francisco, pp. II-157-160.

Matsui T. and Furui, S. (1993) "Concatenated Phoneme Models for Text-Variable Speaker Recognition," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Minneapolis, pp. II-391-394.

Matsui T. and Furui, S. (1994a) "Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability," Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp. 59-62.

Matsui T. and Furui, S. (1994b) "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Adelaide, 13.1.

Matsui T. and Furui, S. (1996a) "Robust Methods of Updating Model and A Priori Threshold in Speaker Verification," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Atlanta, pp. I-97-100.

Matsui T. and Furui, S. (1996b) "Speaker Recognition Using HMM Composition in Noisy Environments," Computer Speech and Language, Vol. 10, pp. 107-116.

Montacie, C., et al. (1992) "Cinematic Techniques for Speech Processing: Temporal Decomposition and Multivariate Linear Prediction," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, San Francisco, pp. I-153-156.

Naik, J., Netsch, M. and Doddington, G. (1989). "Speaker verification over Long Distance Telephone Lines," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S10b.3, pp. 524-527.

Newman, M., Gillick, L., Ito, Y., McAllaster, D. and Peskin, B. (1996) "Speaker Verification through

Large Vocabulary Continuous Speech Recognition," Proc. Int. Conf. Spoken Language Processing, Philadelphia, pp. 2419-2422.

Poritz, A. B. (1982) "Linear Predictive Hidden Markov Models and the Speech Signal," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S11.5, pp. 1291-1294.

Reynolds, D. (1994) "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pp.27-30.

Rose, R. C. and Reynolds, R. A. (1990) "Text Independent Speaker Identification Using Automatic Acoustic Segmentation," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S51.10, pp. 293-296.

Rose, R. C. Hofstetter, E. M. and Reynolds, D. A. (1994) "Integrated Models of Signal and Background with Application to Speaker Identification in Noise," IEEE Trans. Speech and Audio Processing, Vol. 2, No. 2, pp. 245-257.

Rosenberg, A. E. and Soong, F. K. (1987) "Evaluation of a Vector Quantization Talker Recognition System in Text Independent and Text Dependent Modes," Computer Speech and Language, 22, pp. 143-157.

Rosenberg, A. E., Lee, C. -H. and Gokcen, S. (1991) "Connected Word Talker Verification Using Whole Word Hidden Markov Models," Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Toronto, S6.4, pp. 381-384.

Rosenberg, A. E. and Soong, F. K. (1991) "Recent Research in Automatic Speaker Recognition," in Furui, S. and Sondhi M. M.(Eds), *Advances in Speech Signal Processing,* New York: Marcel Dekker, pp. 701-737.

Rosenberg, A. E., DeLong, J., Lee, C-H., Juang, B-H. and Soong, F. K. (1992) "The Use of Cohort Normalized Scores for Speaker Verification," Proc. Int. Conf. Spoken Language Processing, Banff, Th.sAM.4.2, pp. 599-602.

Savic, M. and Gupta, S. K. (1990) "Variable Parameter Speaker Verification System Based on Hidden Markov Modeling," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.7, pp. 281-284.

Siu, M., Yu, G. and Gish, H. (1992) "An Unsupervised, Sequential Learning Algorithm for the Segmentation of Speech Waveforms with Multiple Speakers," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, ppI-189-192.

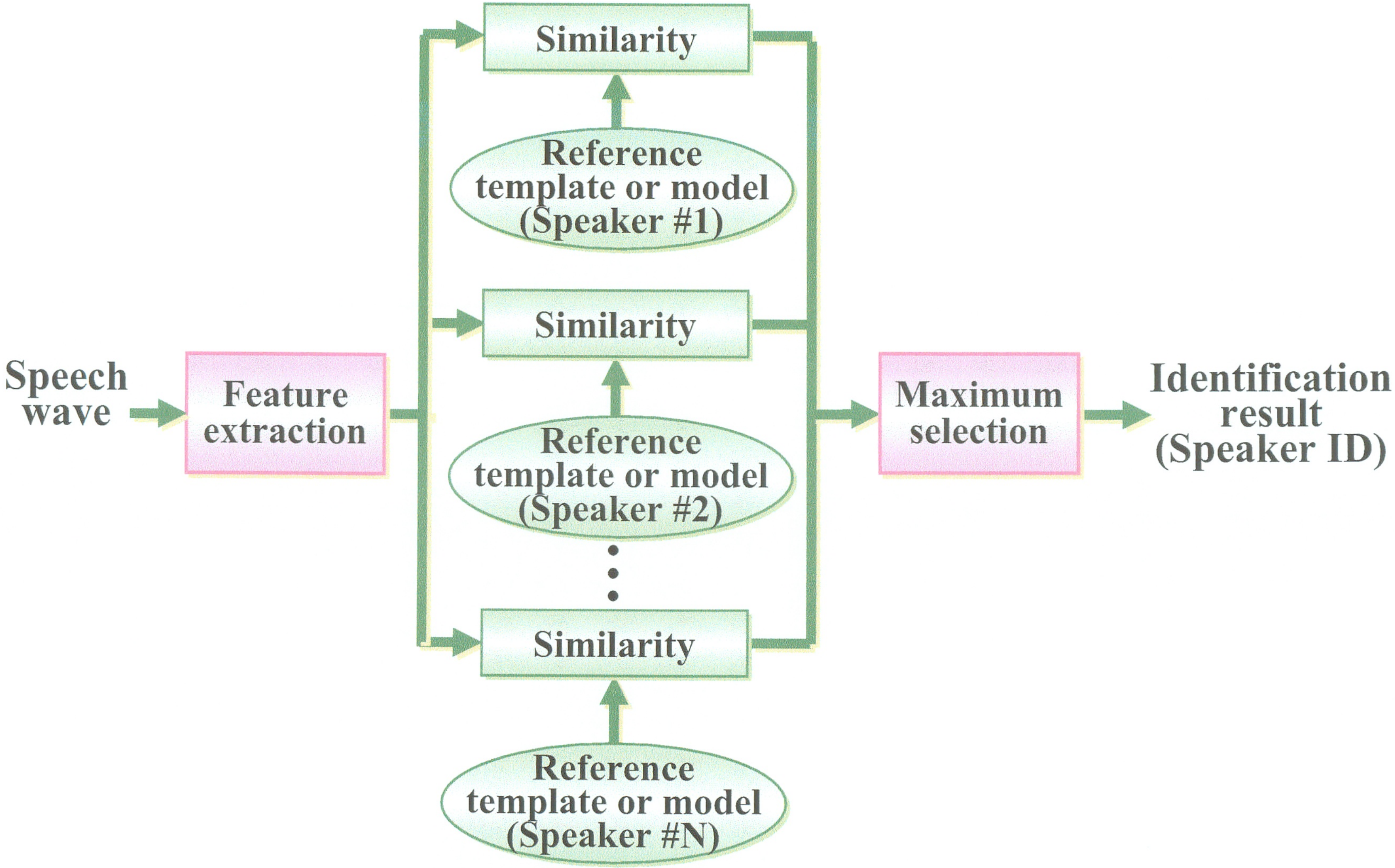
Soong, F. K. and Rosenberg, A. E. (1988)"On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-36, No. 6, pp. 871-879.

Wilcox, L., Chen, F., Kimber, D. and Balasubramanian, V. (1994) "Segmentation of Speech Using Speaker Identification," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. I-161-164.

Figure captions

Fig. 1 – Basic structures of speaker recognition systems.

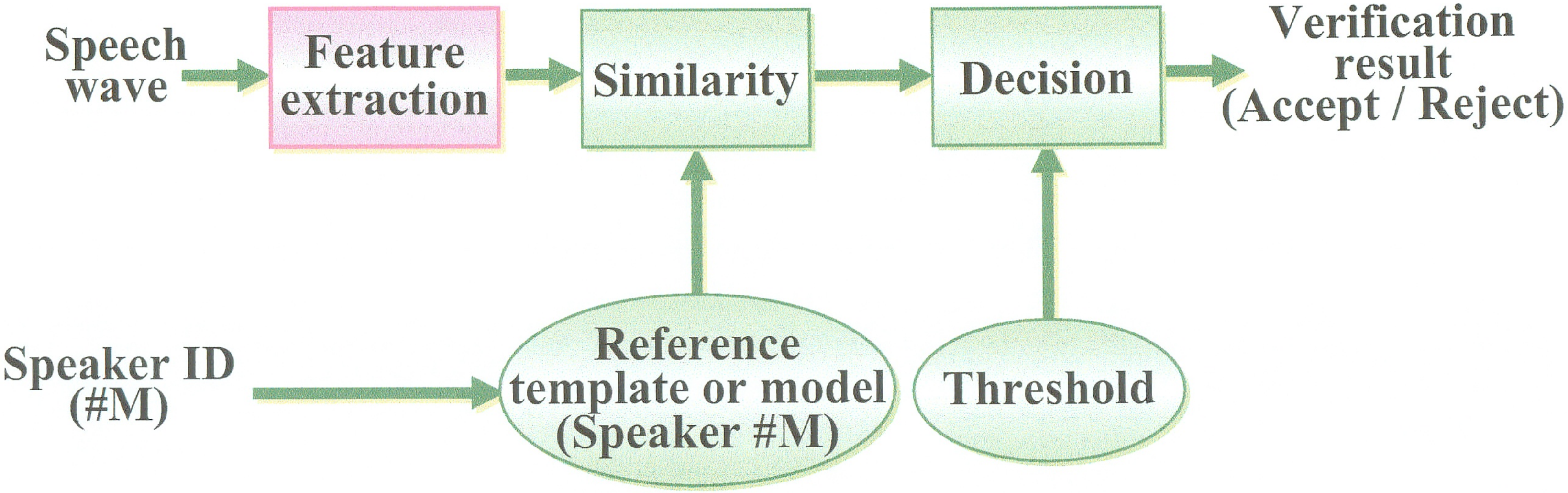
Fig. 2 - Basic structures of text-independent speaker recognition methods.



Speech wave
Feature extraction
Similarity
Reference template or model (Speaker #1)
Similarity
Reference template or model (Speaker #2)
Similarity
Reference template or model (Speaker #N)
Maximum selection
Identification result (Speaker ID)

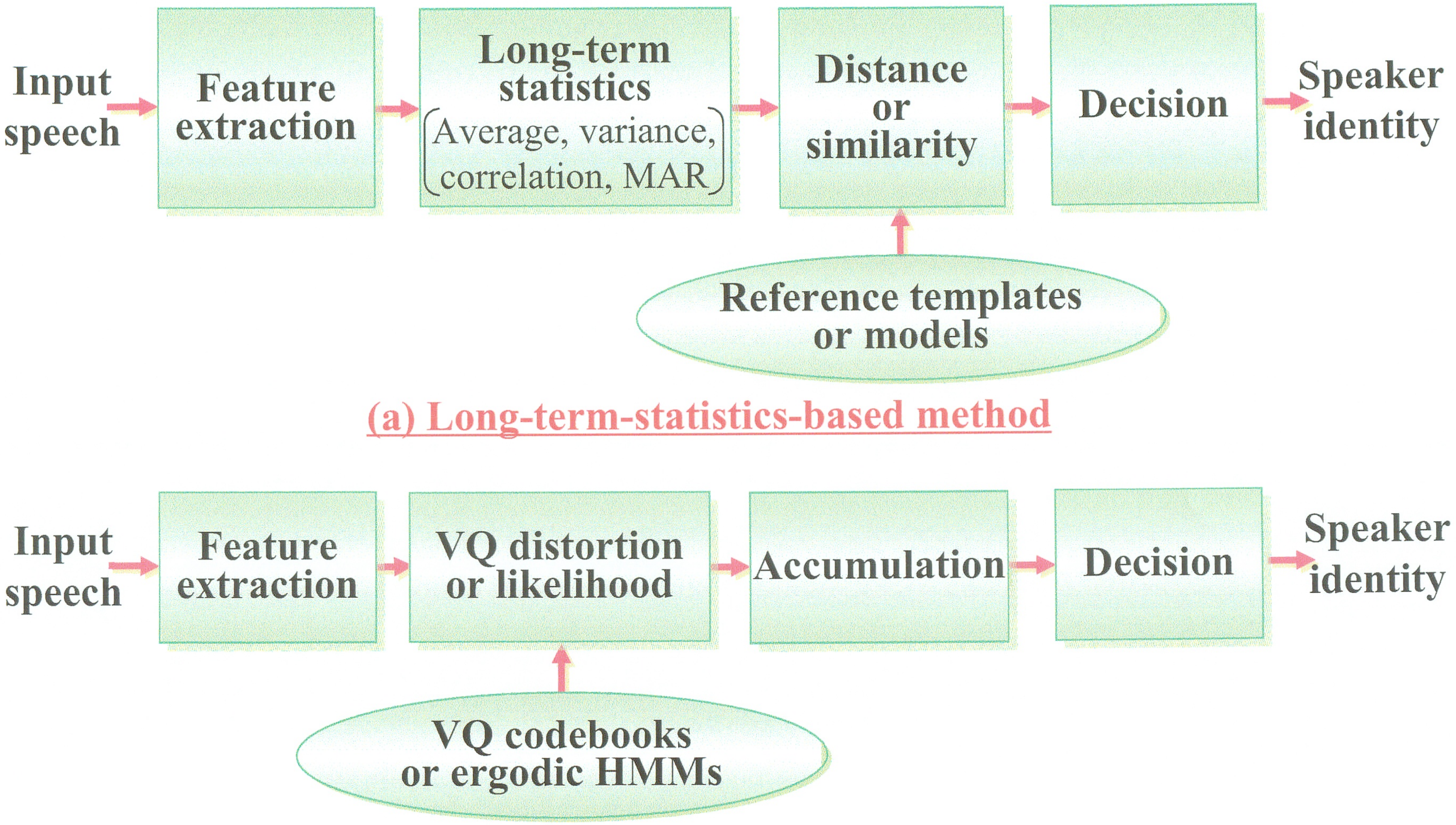
(a) Speaker identification

Fig. 1 - Basic structure of speaker recognition systems.


Speech wave
Feature extraction
Similarity
Decision
Verification result (Accept / Reject)
Speaker ID (#M)
Reference template or model (Speaker #M)
Threshold

(b) Speaker verification

Fig. 1 - Basic structure of speaker recognition systems. (cont.)



Input speech

Feature extraction

Long-term statistics (Average, variance, correlation, MAR)

Distance or similarity

Decision

Speaker identity

Reference templates or models

(a) Long-term-statistics-based method

Input speech

Feature extraction

VQ distortion or likelihood

Accumulation

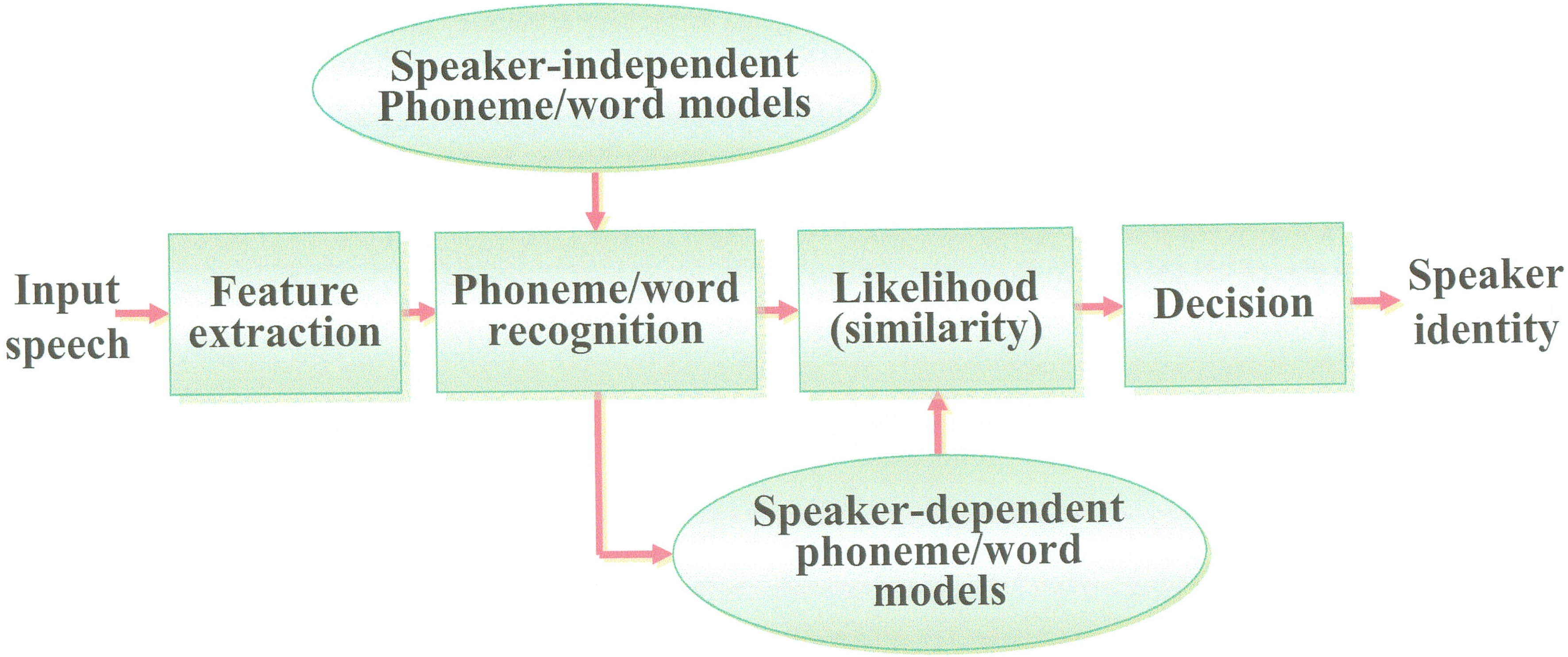
Decision

Speaker identity

VQ codebooks or ergodic HMMs

(b) VQ/HMM-based method

Fig. 2 - Basic structures of text-independent speaker recognition methods.



(c) Speech-recognition-based method

Fig. 2 - Basic structures of text-independent speaker recognition methods. (cont.)