T2R2 東京科学大学 リサーチリポジトリ
Science Tokyo Research Repository

論文 / 著書情報
Article / Book Information

Title	An Overview of Speaker Recognition Technology
Authors	Sadaoki Furui
Citation	ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, , , pp. 1-9
Pub. date	1994,

Powered by T2R2 (Science Tokyo Research Repository)

ESCA WORKSHOP ON AUTOMATIC SPEAKER RECOGNITION, IDENTIFICATION AND VERIFICATION , Martigny, pp. 1-9, (1994)

AN OVERVIEW OF SPEAKER RECOGNITION TECHNOLOGY

Sadaoki Furui

Abstract—**This paper overviews recent advances in speaker recognition technology. The first part of the paper discusses general topics and issues. Speaker recognition can be divided into speaker identification and verification, and into text-dependent and text-independent methods. The second part of the paper is devoted to discussion of more specific topics of recent interest which have led to interesting new approaches and techniques. They include parameter/distance normalization techniques, VQ-/ergodic-HMM-based text-independent recognition methods, and a text-prompted recognition method. The paper concludes with a short discussion assessing the current status and possibilities for the future.**

Keywords— **text-dependent, text-independent, distance normalization, VQ-based method, HMM-based method, text-prompted method.**

1. Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique will make it possible to verify the identity of persons accessing systems, that is, access control by voice, in various services. These services include banking transactions over a telephone network, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Speaker recognition technology is, as such, expected to create new services and make our daily lives more convenient.

This paper is not intended to be a comprehensive review of speaker recognition technology. Rather, it is intended to give an overview of recent advances and the problems which must be solved in the future. The reader is referred to papers by Doddington, Furui, O'Shaugnessy, and Rosenberg and Soong for more general reviews [Doddington, 1985; Furui, 1986, 1989, 1991b; O'Shaugnessy, 1986; Rosenberg and Soong, 1991].

2. Principles of Speaker Recognition

2.1 Classification of Speaker Recognition Technology

Speaker recognition can be divided 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 claim of a speaker. Most of the applications in which voice is used as a key to confirm the identity claim of a speaker are classified as speaker verification. 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 two decision alternatives, accept or reject, regardless of the population size. Therefore, speaker identification performance decreases as the size of population increases, whereas speaker verification performance approaches a constant, independent of the size of population, unless the distribution of physical characteristics of speakers is extremely biased.

Speaker recognition methods can also be divided into text-dependent and text-independent methods. The former require the speaker to provide utterances of the key words or sentences having the same text for both training and recognition trials, whereas the latter do not rely on a specific text being spoken. The text-dependent methods are usually based on template matching techniques in which the time axes of an input speech sample and each reference template or reference model of registered speakers are aligned, and the similarity between them accumulated from the beginning to the end of the utterance is calculated. The structure of text-dependent recognition systems is, therefore, rather simple. Since this method can directly exploit the voice individuality associated with each phoneme or syllable, it generally achieves higher recognition performance than the text-independent method.

However, there are several applications in which predetermined key words cannot be used. In addition, human beings can recognize speakers irrespective of the content of the utterance. Therefore, text-independent methods have recently been actively investigated. Another advantage of text-independent recognition is that it can be done sequentially, until a desired significance level is reached, without the annoyance of repeating the key words again and again.

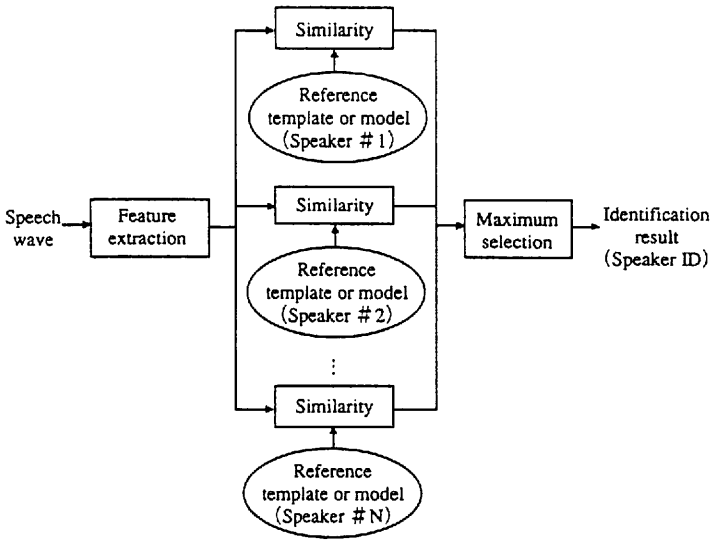
Both text-dependent and independent methods have a serious problem. That is, these systems can easily be defeated, because someone who plays back the recorded voice of a registered speaker uttering key words or sentences into the microphone can be accepted as the registered speaker. To cope with this problem, there are methods in which a small set of words, such as digits, are used as key words and each user is prompted to utter a given sequence of key words that is randomly chosen every time the system is used [Rosenberg et al., 1991; Higgins et al., 1991]. Yet even this method is not reliable enough, since it can be defeated with advanced electronic recording equipment that can reproduce key words in a requested order. Therefore, a text-prompted speaker recognition method has recently been proposed. (See Chaper 6 of this paper.)

S. Furui is with NTT Human Interface Laboratories, 3-9-11, Midoricho, Musashino-shi, Tokyo, 180 Japan, Tel: +81422-59-3910, Fax: +81422-60-7808, E-mail: furui@speech-sun.ntt.jp.

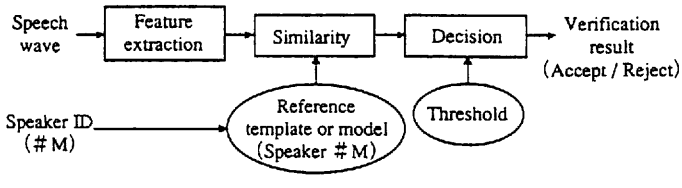
2.2 Basic Structures of Speaker Recognition Systems

Figure 1 shows the basic structures of speaker recognition systems. In speaker identification, a speech utterance from an unknown speaker is analyzed and compared with models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. In speaker verification, an identity claim is made by an unknown speaker, and an utterance of the unknown speaker is compared with the model for the speaker whose identity is claimed. If the match is above a certain threshold, the identity claim is verified. A high threshold makes it difficult for impostors to be accepted by the system, but at the risk of rejecting the genuine person. Conversely, a low threshold ensures that the genuine person is accepted consistently, but at the risk of accepting impostors. In order to set a threshold at a desired level of user rejection and impostor acceptance, it is necessary to know the distribution of customer and impostor scores.

The effectiveness of speaker-verification systems can be evaluated by using the receiver operating characteristics (ROC) curve adopted from psychophysics. The ROC curve is obtained by assigning two probabilities, the probability of correct acceptance and the probability of incorrect acceptance, to the vertical and horizontal axes respectively, and varying the decision threshold [Furui, 1989].



(a) Speaker identification



(b) Speaker verification

Fig. 1. Basic structures of speaker recognition systems.

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

3. Feature Parameters and Normalization Techniques

3.1 Feature Parameters

Speaker identity is correlated with the physiological and behavioral characteristics of the speaker. These characteristics exist both in the spectral envelope (vocal tract characteristics) and in the supra-segmental features (voice source characteristics) of speech. Although it is impossible to separate these kinds of characteristics, and many voice characteristics are difficult to measure explicitly, many characteristics are captured implicitly by various signal measurements. Such signal measurements as short term and long term spectra, and overall energy are easy to obtain. These measurements provide the means for effectively discriminating among speakers. Fundamental frequency can also be used to recognize speakers if it can be extracted reliably [Atal, 1972; Matsui and Furui, 1990].

The current most commonly used short-term spectral measurements are LPC-derived cepstral coefficients and their regression coefficients [Sagayama and Itakura, 1979; Furui, 1979, 1981]. A spectral envelope reconstructed from a truncated set of cepstral coefficients is much smoother than one reconstructed from LPC coefficients, and, therefore, provides a stabler representation from one repetition to another of a particular speaker's utterances. 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- and delta-delta-cepstral coefficients.

3.2 Normalization Techniques

The most significant factor affecting automatic speaker recognition performance is variation in signal characteristics from trial to trial (intersession variability, 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 tokens 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. Two types of normalization techniques have been tried; one in the parameter domain, and the other in the distance/similarity domain.

3.2.1 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, 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. Additive variation in the log spectral domain can be fairly well compensated by this method. This method, however, unavoidably removes some text-dependent and speaker specific features, and, therefore, is inappropriate for short utterances in speaker recognition applications.

Gish [1990] demonstrated that by simply prefiltering the speech transmitted over different telephone lines with a fixed filter, text-independent speaker recognition performance can be significantly improved. Gish et al. [1985, 1986] have also proposed using multi-variate Gaussian probability density functions to model channels statistically. This can be achieved if enough training samples of channels to be modeled are available. It was shown that time derivatives of cepstral coefficients (delta-cepstral coefficients) are resistant to linear channel mismatch between training and testing [Soong and Rosenberg, 1988].

3.2.2 Distance/Similarity-Domain Normalization

Higgins et al. [1991] proposed a normalization method for distance (similarity, likelihood) values that uses a likelihood ratio. The likelihood ratio is defined as the ratio of the conditional probability of the observed measurements of the utterance given the claimed identity to the conditional probability of the observed measurements given the speaker is an imposter. A mathematical expression for the likelihood ratio is

$$\log L(X) = \log p(X|S = S_c) - \log p(X|S \neq S_c) \quad (1)$$

Generally, a positive value of $\log L$ indicates a valid claim, whereas a negative value indicates an imposter. We call the second term of the right hand side of the Eq. (1) the normalization term.

The density at point X for all speakers other than true speaker S can be dominated by the density for the nearest reference speaker, if we assume that the set of reference speakers is representative of all speakers. We can therefore arrive at the decision criterion

$$\log L(X) = \log p(X|S = S_c) - \max_{S \in Ref, S \neq S_c} \log p(X|S) \quad (2)$$

This shows that likelihood ratio normalization approximates optimal scoring in Bayes' sense.

There are basically two possible sets of speakers, "cohort speakers" to be chosen for calculating the normalization term of the Eq. (1): the speakers that are typical of the general population and the speakers that are representative of the population near the claimed speaker. Higgins et al. proposed use of the latter set, which is expected to increase the selectivity of the algorithm against voices similar to the claimed speaker:

$$\log L(X) = \log p(X|S = S_c) - \log \sum_{S \in Cohort, S \neq S_c} p(X|S) \quad (3)$$

Experimental results show that this normalization method improves speaker separability and reduces the need for speaker-dependent or text-dependent thresholding, compared with scoring using only the model of the claimed speaker. Another experiment in which the size of the cohort speaker set was varied from 1 to 5 showed that speaker verification performance increases as a function of the cohort size, and that the use of normalization significantly compensates for the degradation obtained by comparing verification utterances recorded using an electret microphone with models constructed from training utterances recorded with a carbon button microphone [Rosenberg, 1992].

Matsui and Furui [1993] proposed a normalization method based on a posteriori probability:

$$\log L(X) = \log p(X|S = S_c) - \log \sum_{S \in Ref} p(X|S) \quad (4)$$

The difference between the normalization method based on the likelihood ratio and that based on a posteriori probability is in whether or not the claimed speaker is included in the speaker set for normalization; the cohort set or the speaker set in the likelihood-ratio-based method does not include the claimed speaker, whereas the normalization term for a posteriori-probability-based method is calculated by using all the reference speakers, including the claimed speaker. Matsui et al. approximated the summation in Eq. (4) by the summation over a small set of speakers having relatively high likelihood values. Experimental results indicate that the two normalization methods are almost equally effective.

Matsui and Furui [1994a] recently proposed a new method in which the normalization term is approximated by the likelihood for a Gaussian mixture which models the parameter distribution for free-text utterances by all the reference speakers. This method has been confirmed to give much better results than either of the above-mentioned normalization methods.

Since these normalization methods neglect the absolute deviation between the claimed speaker's model and the input speech, they cannot differentiate highly dissimilar speakers. 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.

4. Text-Dependent Speaker Recognition Methods

4.1 DTW-Based Methods

A typical approach to text-dependent speaker recognition is the spectral template matching approach. 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 dynamic programming time warping (DTW) algorithm.

Figure 2 shows an example of a typical structure of the DTW-based system [Furui, 1981]. Initially, 10 LPC cepstral coefficients are extracted every 10 ms from a short sentence of speech. The spectral equalization technique described in the previous section is applied to each cepstral coefficient to compensate for transmission distortion and intraspeaker variability. In addition to the normalized cepstral coefficients, delta and delta-delta cepstral coefficients are extracted every 10 ms. 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.

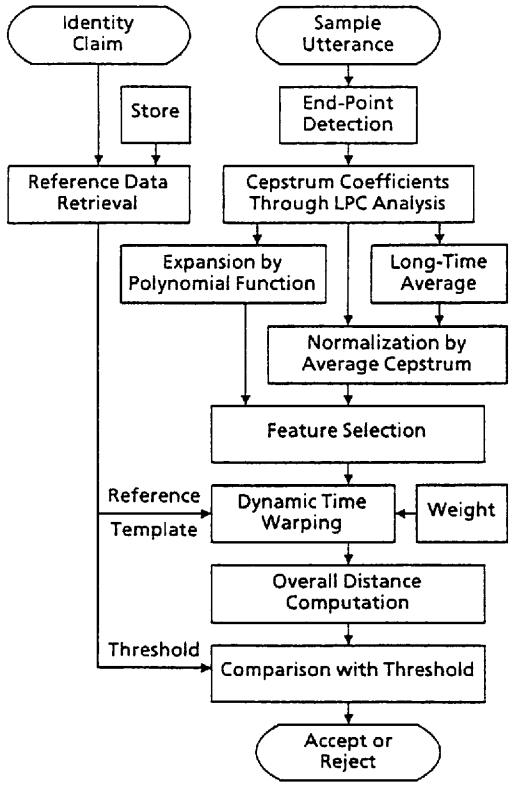


Fig. 2. A typical structure of the DTW-based system.

4.2 HMM-Based Methods

HMM (the hidden Markov model) has the capability of efficiently modeling statistical variation in spectral features. Therefore, HMM-based methods have achieved significantly better recognition accuracies than the DTW-based methods [Zheng and Yuan, 1988; Naik et al., 1989; Rosenberg et al., 1991].

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. The results of experiments using isolated digit utterances show only small differences in performance between PLU- and ASU-based representations.

5. Text-Independent Speaker Recognition Methods

In text-independent speaker recognition, the words or sentences used in recognition trials cannot generally be predicted. Since it is impossible to model or match speech events at the word or sentence level, the following five kinds of methods have been investigated.

5.1 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 [Furui et al., 1972; Markel et al., 1977; Markel and Davi, 1979]. However, long-term spectral averages are extreme condensations of the spectral characteristics of a speaker's utterances and, as such, lack the discriminating power included in 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 is reduced by introducing a weighted cepstral distance measure.

Studies on using statistical dynamic features have also been reported. Montacie et al. [1992] used a multivariate auto-regression (MAR) model to characterize speakers, and reported good speaker recognition results. Griffin et al. [1994] studied distance measures for the MAR-based method, and reported that when 10 sentences were used for training and one sentence was used for testing, identification and verification rates were almost the same as obtained by an HMM-based method. In these experiments, the MAR model was applied to the time series of cepstral vectors. It was also 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.

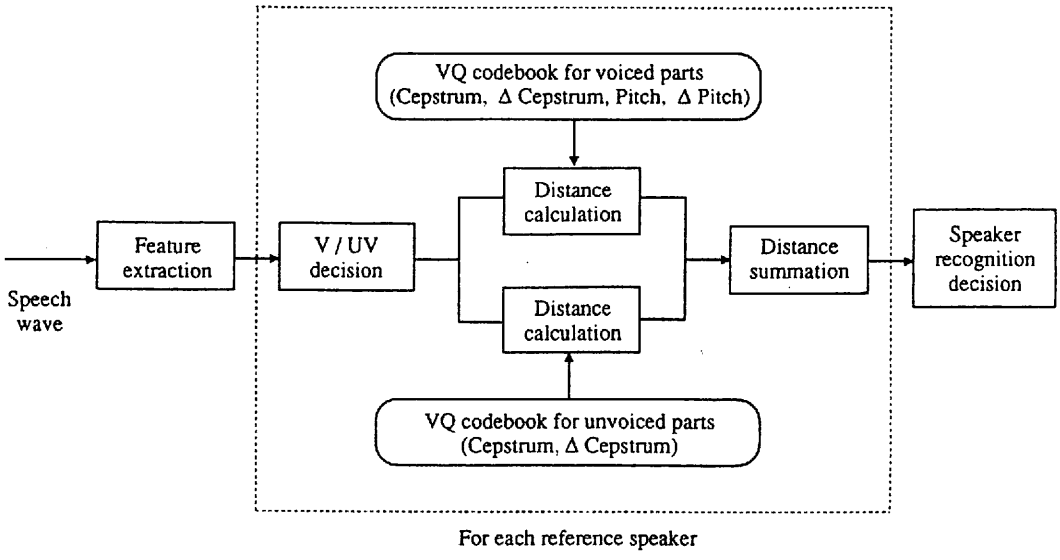


Fig. 3. A structure of the VQ-based method using feature vectors consisting of instantaneous and transitional features calculated for both cepstral coefficients and fundamental frequency.

5.2 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, efficient ways of compressing the training data have been tried 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, Jr., 1983; Shikano, 1985; Soong et al., 1987; Rosenberg and Soong, 1987; Matsui and Furui, 1990, 1991]. A speaker-specific codebook is generated by clustering the training feature vectors of each speaker. In the recognition stage, an input utterance is vector-quantized using the codebook of each reference speaker, and the VQ distortion accumulated over the entire input utterance is used in making the recognition decision.

Figure 3 shows a method using a codebook for long feature vectors consisting of instantaneous and transitional features calculated for both cepstral coefficients and fundamental frequency [Matsui and Furui, 1990, 1991]. 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 is introduced to take into account the intra- and inter-speaker variability and to deal with the outlier problem in the distribution of feature vectors. The outlier vectors correspond to intersession spectral variation and the difference between phonetic content of the training texts and the test utterances. Experimental results confirmed high recognition accuracies even when the codebooks for each speaker were made using training utterances recorded in a single session and the time difference between training and testing was more than three months. It was also confirmed that, although fundamental frequency achieved only a low recognition rate by itself, the recognition accuracy was largely improved by combining 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 [Sugiyama, 1988; 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 [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 capability for handling temporal variation.

5.3 Ergodic-HMM-Based Methods

On a long time scale, temporal variation in speech signal parameters can be represented by stochastic Markovian transitions between states. 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. Poritz characterized the automatically obtained categories as strong voicing, silence, nasal/liquid, stop burst/post silence, and frication.

Savic and Gupta [1990] also used a five-state ergodic lin-

ear predictive HMM for broad phonetic categorization. After identifying frames which belong to particular phonetic categories, a feature selection was performed. In the training phase, reference templates are generated and verification thresholds are computed for each phonetic category. In the verification phase, after the 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 were chosen to reflect the effectiveness of particular categories of phonemes in discriminating between speakers and are adjusted to maximize the verification performance. Experimental results show that verification accuracy can be considerably improved by this category-dependent weighted linear combination method.

Tishby [1991] extended Poritz's work to the richer class of mixture autoregressive (AR) HMMs. In these models, the states are described as a linear combination (mixture) of AR sources. It can be shown that mixture models are equivalent to a larger HMM with simple states, together with additional constraints on the possible transitions between states.

The performance of the speaker recognition method using codebooks representing both cepstral and pitch characteristics, described above, has been improved by introducing an ergodic HMM for broad phonetic categorization [Matsui and Furui, 1992]. In that approach, a VQ-based method and discrete/continuous ergodic HMM-based methods are compared, in particular from the viewpoint of robustness against utterance variations. It was shown that a continuous ergodic HMM method is far superior to a discrete ergodic HMM method, and that a continuous ergodic HMM method is as robust as a VQ-based method when enough training data is available. However, when little data is available, the VQ-based method is more robust than a continuous HMM method. It was also shown that the information on transitions between different states is ineffective for text-independent speaker recognition, and, therefore, the speaker recognition rates using a continuous ergodic HMM are strongly correlated with the total number of mixtures, irrespective of the number of states. Figure 4 shows speaker identification rates as a function of the number of states and mixtures.

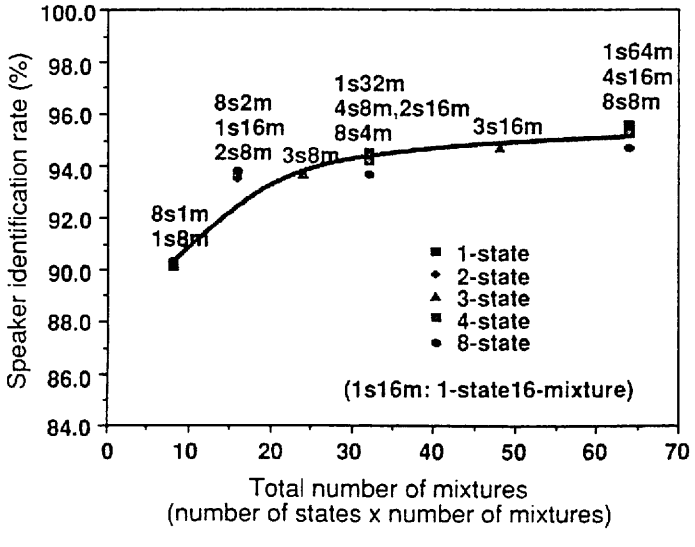


Fig. 4. Speaker identification rates as a function of the number of states and mixtures in ergodic HMMs.

Rose and Reynolds [1990] investigated a technique based on maximum likelihood estimation of a Gaussian mixture model representation of speaker identity. This method corresponds to the single-state continuous ergodic HMM investigated by Matsui et al. Furthermore, a VQ-based method can be viewed as a special (degenerate) case of a single-state HMM with a distortion measure used as the observation probability. Gaussian mixtures are noted for their robustness as a parametric model and their ability to form smooth estimates of rather arbitrary underlying densities. Broad phonetic categorization can also be implemented by a speaker-specific hierarchical classifier instead of an HMM, and the effectiveness of this approach has also been confirmed [Eatock and Mason, 1990].

The ASU-based speaker verification method described in Section 4.2 has also been tested in the text-independent mode [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 Methods

Speaker recognition based on feed-forward neural net models have been investigated [Oglesby and Mason, 1990]. Each registered speaker has a personalized neural net that is trained to be activated only by that speaker's utterances. It is assumed that including speech from many people in the training data of each net enables direct modeling of the differences between the authorized person's speech and an imposter's speech. It has been found that while the net architecture and the amount of training utterances strongly affect the recognition performance, it is comparable to the performance of the VQ approach based on personalized codebooks.

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

5.5 Event-Specific-Characteristics-Based Methods

Many studies have also been carried out to extract and characterize specific events thought to have good speaker discriminating properties. Kao et al. [1992] used a speaker-independent speech recognizer to hypothesize phonetic segments, and adopted speaker-specific VQ codebooks for each phonetic class.

These studies, however, have not resulted in practical recognition systems because spectral and temporal variations make it difficult to reliably segment and label specific speech events across different utterances and speakers. That is, the present technology of the speaker-independent

phonetic typewriter is far from satisfactory; it makes the total speaker recognition system too complicated and is thus far from practical.

6. Text-Prompted Speaker Recognition Method

6.1 Key Idea of the Text-Prompted Method

The most suitable application for speaker recognition techniques is access control. In such applications, users can be prompted to provide an identity claim as well as utterances of specific key words or sentences. In the text-prompted speaker recognition method [Matsui and Furui, 1993, 1994b], the recognition system prompts each user with a new key sentence every time the system is used, and accepts the input utterance only when it decides that the registered speaker has uttered the prompted sentence. The sentence can be displayed as characters or spoken by a synthesized voice. Because the vocabulary is unlimited, prospective impostors cannot know in advance the sentence they will be prompted to say. This method not only can accurately recognize speakers but also can reject utterances whose text differs from the prompted text, even if it is uttered by the registered speaker. A recorded voice can thus be correctly rejected.

6.2 System Structure

Figure 5 shows a block diagram of the method. This method is facilitated by using 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 with 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. Since the text of training utterances is known, these utterances can be modeled as the concatenation of phoneme models, and these models can be automatically adapted by an iterative algorithm. In order to properly adapt the models of phonemes that are not included in the training utterances, a new adaptation method based on tied-mixture HMMs has recently been proposed [Matsui and Furui, 1994b].

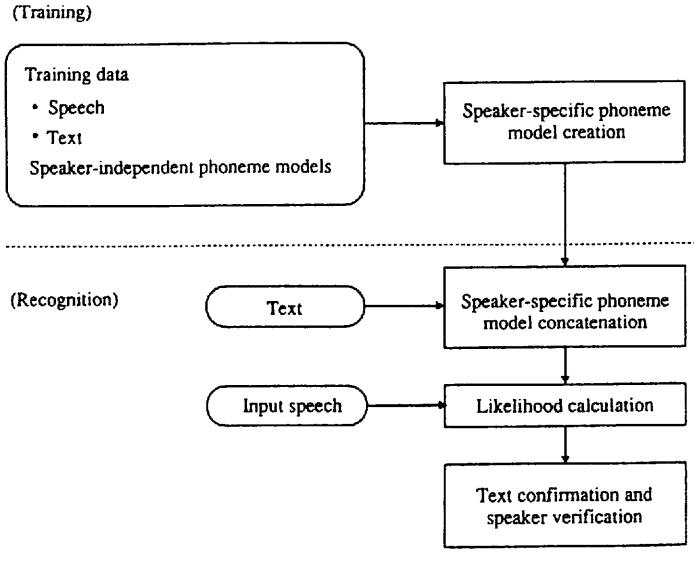


Fig. 5. Block diagram of the text-prompted speaker recognition method.

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 recognition decision. If the likelihood of both speaker and text is high enough, the speaker is accepted as the claimed speaker.

As described in Section 3.2, how to accommodate speech variation is important in speaker recognition. Especially in the case of text-prompted speaker recognition, where speech with different texts are uttered at different sessions, the likelihood has a wide range. The likelihood normalization based on likelihood ratio or a posteriori probability is, therefore, indispensable in setting a stable threshold for speaker and text verification.

6.3 Recognition Experiments

Recognition experiments were performed to evaluate the effectiveness of this method. Various sentences uttered by 15 speakers (10 male and 5 female) at three sessions over a period of roughly five months were used. The results show that, when the adaptation method for tied-mixture-based phoneme models and the likelihood normalization method were used, a speaker and text verification rate of 99.4% was obtained.

7. Relationship to Speech Recognition Technology

Recently, speaker-independent speech recognition methods using HMM techniques have been actively investigated, and the recognition accuracy has been largely improved. However, one of the disadvantages of the speaker-independent approach is that it neglects various useful characteristics of the speaker, and, therefore, speaker-independent recognition methods can hardly reach the accuracies achieved by speaker-dependent methods. When the distributions 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.

If speaker-specific characteristics can be properly used, the recognition process is expected to be accelerated due to the narrowing of the search space, and higher recognition accuracies will be obtained. In order to do this, it is essential to introduce speaker adaptation techniques [Furui, 1991a].

Speaker recognition and speaker adaptation research have long been conducted separately, since it has not necessarily been realistic to use common techniques to achieve best performances in both areas. However, in the case of text-prompted speaker recognition, it is crucially important to create speaker-specific phoneme models that bear

enough information related to both each phoneme and the speaker. An interesting research topic is the automatic adjustment of speaker-independent phoneme models to each new speaker so that the performance of both speech and speaker recognition are simultaneously improved. Speaker adaptation techniques will, therefore, be investigated using a common approach, and become a core part of both speaker and speech recognition algorithms.

8. Future Problems

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over a long period, insensitive to the variation of speaking manner, including speaking rate and level, and robust against the variation of voice quality such as those due to voice disguise or colds. It is also important to develop a method to cope with the problems of distortion due to telephone sets and channels, and background and channel noises.

Recent advances in speaker recognition are mainly due to improvements in techniques for making speaker-sensitive feature measures and models, and they have not necessarily come about as an outgrowth of new or better understanding of speaker characteristics or how to extract them from the speech signal. It can be expected that better understanding of speaker characteristics in the speech signal can be applied to provide more effective speaker recognition systems.

As fundamental research, it is important to pursue a method for extracting and representing the speaker characteristics that are commonly included in all the phonemes irrespective of the speech text.

From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled.

Studies on automatic extraction of the speech periods of each person separately from a dialogue involving more than two people have recently appeared as an extension of speaker recognition technology [Gish et al., 1991; Siu et al., 1992].

Speaker characterization techniques are also related to the research on improving synthesized speech quality by adding natural characteristics of voice individuality, and converting synthesized voice individuality from one speaker to another [Abe et al., 1988]. It is expected that diversified research related to speaker-specific information in speech waves will become more active in the near future.

References

M. Abe, S. Nakamura, K. Shikano and H. Kuwabara (1988), *"Voice Conversion through Vector Quantization,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S14.1, pp. 655-658.

B. S. Atal (1972), *"Automatic Speaker Recognition Based on Pitch Contours,"* J. Acoust. Soc. Am., Vol. 52, No. 6, pp. 1687-1697.

B. S. Atal (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.

Y. Bennani, F. Fogelman Soulie and P. Gallinari (1990), *"A Connectionist Approach for Automatic Speaker Identification,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.2, pp. 265-268.

G. R. Doddington (1985), *"Speaker Recognition-Identifying People by their Voices,"* Proc. IEEE, Vol. 73, No. 11, pp. 1651-1664.

J. Eatock and J. S. Mason (1990), *"Automatically Focusing on Good Discriminating Speech Segments in Speaker Recognition,"* Proc. Int. Conf. Spoken Language Processing, 5.2, pp. 133-136.

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

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 (1979), *"New Techniques for Automatic Speaker Verification Using Telephone Speech,"* J. Acoust. Soc. Am. (abstract), Suppl. 1, No. 66, p. S35.

S. Furui (1981), *"Cepstral Analysis Technique for Automatic Speaker Verification,"* IEEE Trans. Acoust. Speech Signal Processing, 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 (1989), *"Digital Speech Processing, Synthesis, and Recognition,"* Marcel Dekker, New York.

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

S. Furui (1991b), *"Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques,"* Speech Communication, Vol. 10, No. 5-6, pp.505-520.

H. Gish, M. Krasner, W. Russell and J. Wolf (1986), *"Methods and Experiments for Text-Independent Speaker Recognition over Telephone Channels,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, 17.2, pp. 865-8.

H. Gish (1990), *"Robust Discrimination in Automatic Speaker Identification,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.9, pp. 289-292.

H. Gish, K. Karnofsky, K. Krasner, S. Roucos, R. Schwartz and J. Wolf, (1985), *"Investigation of Text-Independent Speaker Identification over Telephone Channels,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 379-382.

H. Gish, M. -H. Siu and R. Rohlicek (1991), *"Segregation of Speakers for Speech Recognition and Speaker Identification,"* Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, Toronto, S13.11, pp. 873-876.

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

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

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

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

Y. -H. Kao, P. K. Rajasekaran and J. S. Baras (1992), *"Free-Text Speaker Identification over Long Distance Telephone Channel Using Hypothesized Phonetic Segmentation,"* Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing, San Francisco, pp. II-177-180.

K. -P. Li and E. H. Wrench Jr. (1983), *"An Approach to Text-Independent Speaker Recognition with Short Utterances,"* Proc.

IEEE Int. Conf. Acoust., Speech, Signal Processing, 12.9, pp. 555-558.

J. D. Markel, B. T. Oshika and A. H. Gray (1977), *"Long-Term Feature Averaging for Speaker Recognition,"* IEEE Trans. Acoust. Speech Signal Processing, Vol. ASSP-25, No. 4, pp. 330-337.

J. D. Markel and S. B. Davi (1979), *"Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base,"* IEEE Trans. Acoust. Speech Signal Processing, Vol. ASSP-27, No. 1, pp. 74-82.

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

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

T. Matsui and S. Furui (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.

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

T. Matsui and S. Furui (1994a), *"A New Similarity Normalization Method for Speaker Verification Based on a Posteriori Probability,"* ESCA Workshop on Automatic Speaker Recognition, Identification and Verification.

T. Matsui and S. Furui (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.

C. Montacie 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.

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

D. O' Shaugnessy (1986), *"Speaker Recognition,"* IEEE ASSP Magazine, 3, No. 4, pp. 4-17.

J. Oglesby and J. S. Mason (1990), *"Optimization of Neural Models for Speaker Identification,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.1, pp. 261-264.

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

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

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

A. E. Rosenberg, C. -H. Lee and F. K. Soong (1990a), *"Sub-Word Unit Talker Verification Using Hidden Markov Models,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S5.3, pp. 269-272.

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. Int. Conf. Spoken Language Processing, 5.4, pp. 141-144.

A. E. Rosenberg, C. -H. Lee and S. Gokcen (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.

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

A. E. Rosenberg (1992), *"The Use of Cohort Normalized Scores for Speaker Verification,"* Proc. Int. Conf. Spoken Language Processing, Banff, Th.sAM.4.2, pp. 599-602.

S. Sagayama and F. Itakura (1979), *"On Individuality in a Dynamic Measure of Speech,"* Proc. Spring Meeting of Acoust. Soc. Japan (in Japanese), pp. 589-590.

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

K. Shikano (1985), *"Text-Independent Speaker Recognition Experiments Using Codebooks in Vector Quantization,"* J. Acoust. Soc. Am. (abstract), Suppl. 1, No. 77, p. S11.

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

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 Processing, Vol. ASSP-36, No. 6, pp. 871-879.

F. K. Soong, A. E. Rosenberg and B. -H. Juang (1987), *"A Vector Quantization Approach to Speaker Recognition,"* AT&T Technical Journal, No. 66, pp. 14-26.

M. Sugiyama (1988), *"Segment Based Text Independent Speaker Recognition,"* Proc. Spring Meeting of Acoust. Soc. Japan (in Japanese), pp. 75-76.

N. Z. Tishby (1991), *"On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition,"* IEEE Trans. Acoust. Speech, Signal Processing, Vol. ASSP-30, No. 3, pp. 563-570.

Y. -C. Zheng and B. -Z. Yuan (1988), *"Text-Dependent Speaker Identification Using Circular Hidden Markov Models,"* Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, S13.3, pp. 580-582.