# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | |
| Title(English) | Speaker Recognition Technology |
| 著者(和文) | 古井 貞熙 |
| Authors(English) | Tomoko Matsui, Sadaoki Furui |
| 出典(和文) | , Vol. 7, No. 2, pp. 40-48 |
| Citation(English) | NTT Review, Vol. 7, No. 2, pp. 40-48 |
| 発行日 / Pub. date | 1995, 3 |

# Speaker Recognition Technology

**Tomoko Matsui and Sadaoki Furui**

*Speaker recognition is the process of automatically recognizing who is speaking. This paper introduces three speaker recognition methods; a text-dependent method which uses predetermined keywords, a text-independent method which does not rely on a specific text being spoken, and a text-prompted method in which the system prompts each user with a new key sentence every time the system is used. Recently, AT&T and TI (with US Sprint) started field tests and real application of speaker recognition technology, and the Sprint's Voice Phone Card has already been used by many customers. However, there still remain many technical problems to be solved. Among them, this paper introduces several issues that have recently been investigated at NTT Laboratories.*

## Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individuality information in speech waves. This technique will make it possible to verify the identity of persons accessing systems. 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.

There are two types of speaker recognition: speaker identification and verification (see Fig. 1). Speaker identification determines from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the claimed identity of a speaker. Most applications of speaker recognition come under speaker verification. The above examples for verifying the identity of persons accessing systems, that is access control by voice, banking transactions, and etc., are applications of speaker verification. Therefore, this paper focuses on speaker verification, studies the basic system structure, technical problems for constructing the system, and applications, and discusses future problems.

## 1 Speaker Verification

Speaker verification can be divided into three types, text-dependent, text-independent and text-prompted (see Table 1). Text-dependent verification requires the speaker to utter key words or sentences using the same text for both training and verification. Text-independent verification does not rely on a specific text being spoken. Since the former

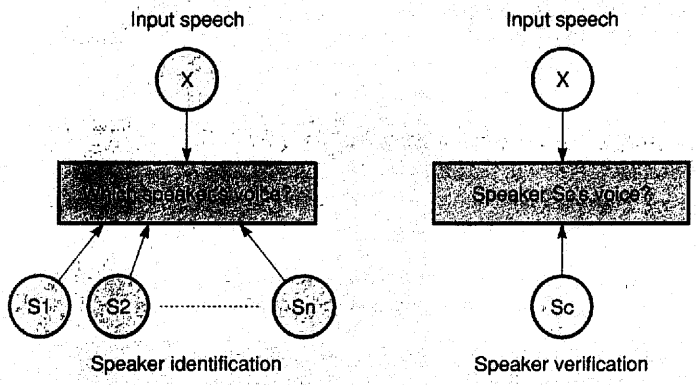


Figure 1. Speaker Identification and Verification

can use text-specific voice information, it can usually obtain higher performance than the latter. Therefore, text-independent verification needs more input speech to obtain similar levels of performance.

Both of these verification systems have a problem that they can be easily fooled simply by playing back the recorded voice of the target speaker. Therefore, we have proposed a text-prompted verification method, in which an arbitrary key sentence is prompted every time. The system verifies the speaker's identity only when it decides the correct prompt has been spoken by the claimed speaker. This verification system is therefore secure against tape-recorders.

The following sections explain about the basic structures of these three types of speaker verification methods.

### *1.1 Text-Dependent Verification*

The basic structure of the text-dependent speaker verification system is shown in Fig. 2. During the training phase, each speaker utters a fixed key text several times. For each utterance, the speech period is detected and analyzed to extract the sequence for each feature parameter (cepstrum, delta-cepstrum, and so on). Then, the sequences are aligned on a time axis by using the dynamic programming (DP) matching technique and averaged to make a reference pattern for the speaker.

Table 1. Speaker Verification Methods

| Method | Utterance | Features and Problems |
|---|---|---|
| Text-Dependent | Fixed keyword or sentence | –High-performance<br>–The system can easily be fooled by tape-recorders |
| Text-Independent | Arbitrary word or sentence | –Long speech is needed<br>–The system can easily be fooled by tape-recorders |
| Text-Prompted | Prompted word or sentence | –The system is secure against tape-recorders |

During the verification phase, the feature parameters are extracted in the same sequences as in the training phase. The time-aligned distance between the sequence and the reference pattern of the claimed speaker is calculated using DP-matching; it is then compared with a predetermined threshold. If the distance is less than the threshold, the input utterance is accepted as the true speaker's, whereas if it is larger, it is rejected.

In the last few years, the hidden Markov model (HMM) has become commonly used instead of the reference pattern for the feature parameter sequences. The likelihood of the HMM for the claimed speaker is calculated and then similarly compared with a threshold to verify speaker identity.

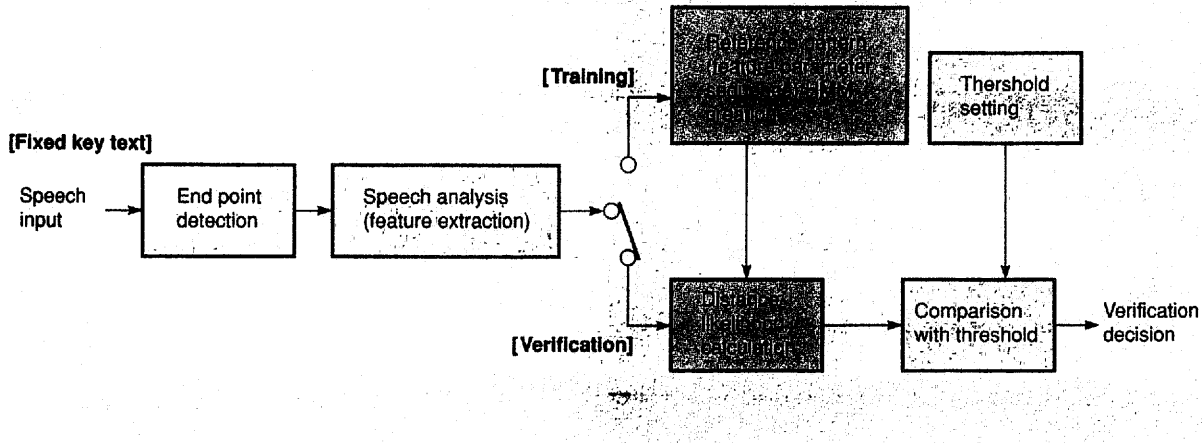


Figure 2. Text-Dependent Speaker Verification System

In text-dependent verification, a reference pattern or model for each speaker is represented by using sequences of feature parameters, which depend on the phoneme sequences in the key text. Therefore, the voice characteristics of each speaker can be reliably extracted from comparatively short utterances. Moreover, a large amount of training data is not necessary because the key text is fixed. Several experiments using telephone lines have shown that this method performs well. In practical use, since a different key text can be used for each speaker, the claimed identity can also be verified by using the key text as a password, thus improving performance even more. (Of course, it is assumed that the key texts cannot be stolen easily.)

### *1.2 Text-Independent Verification*

In text-independent verification, since the text spoken by the user can vary each time, it is impossible to represent a reference pattern (model) for each speaker by using sequences of feature parameters, which depend on the phoneme sequences in a key text. The basic structure of a text-independent speaker verification system is shown in Fig. 3. During the training phase, each speaker utters several sentences in order for the system to obtain the speaker's voice characteristics for various texts or phonemes. The speech period is detected and analyzed to extract the sequence for each feature parameter (cepstrum and so on). Then, vector quantization (VQ) codebooks or ergodic HMMs are made for each speaker from the sequences. In the VQ-based method, the feature vectors are clustered together without paying attention to the time alignment, and mean vectors, or centroids, are calculated for each cluster. The centroids, which represent speaker-specific features, form a "codebook," which can have around 256 centroids, for example. In the HMM-based method, temporal variations in feature parameters can be represented by stochastic Markovian transitions between states (for example, one to five states) and output probabilities for each state. The output probability for each state is approximated by multiple Gaussian distributions (for example, 64 distributions).

During the verification phase, the feature parameters are extracted in the same sequence as in the training phase. In the VQ-based method, the sequence is vector-quantized using the codebook of the claimed speaker; the average distortion is calculated and is compared with the threshold. In the HMM-based method, the likelihood value for the input sequence to match the HMM of the claimed speaker is calculated and compared with the threshold for speaker decision.

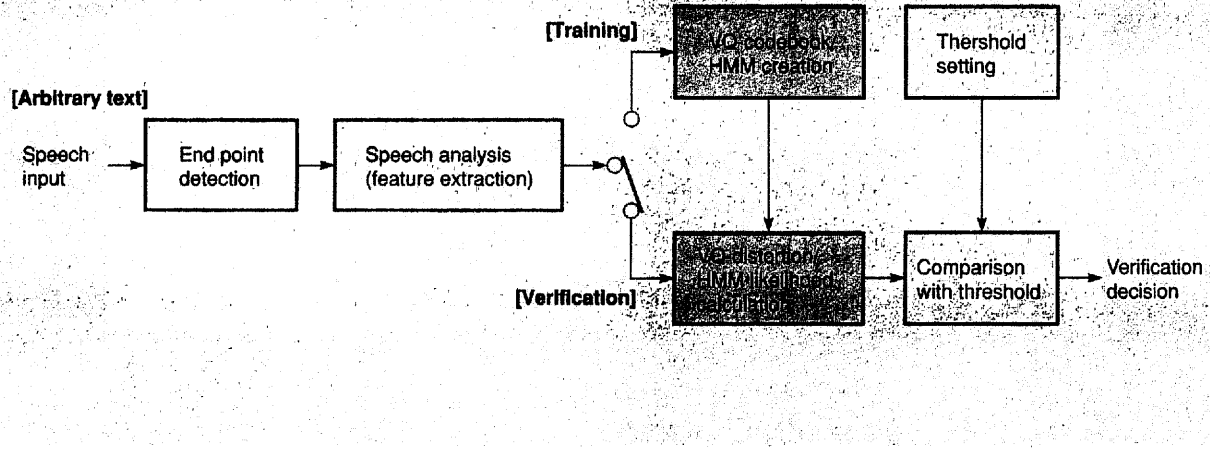


Figure 3. Text-Independent Speaker Verification System

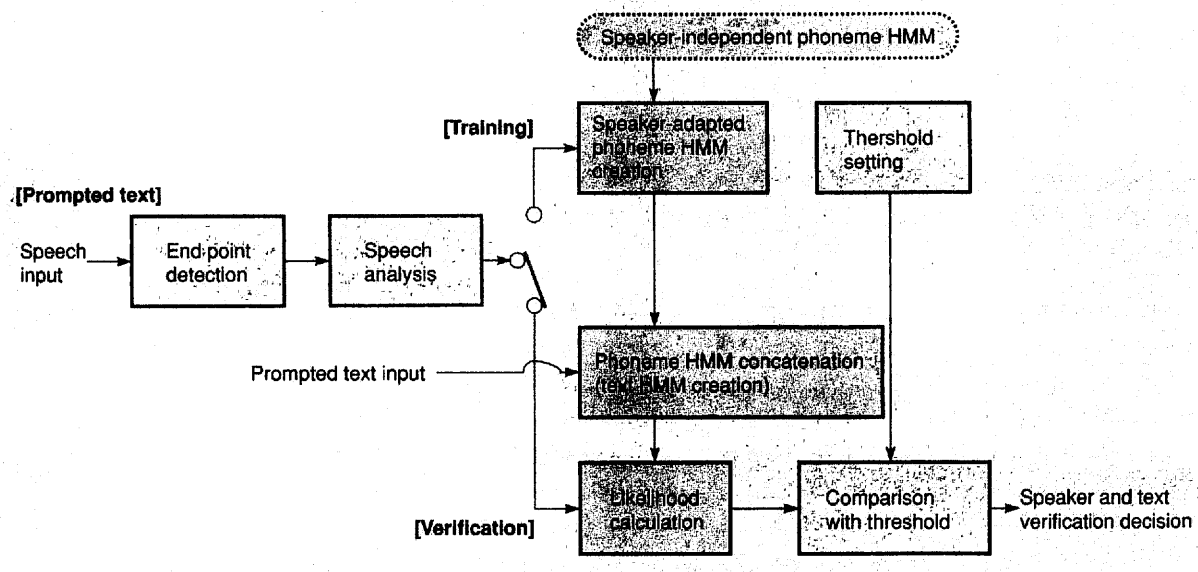


Figure 4. Text-Prompted Speaker Verification System

### *1.3 Text-Prompted Verification*

The basic structure of the text-prompted speaker verification system is shown in Fig. 4. In text-prompted speaker verification, speaker-specific phoneme models are used as basic acoustic units for verifying both the text and the speaker. In order to verify the speaker with a new key text every time the system is used, phoneme models are trained for each speaker and concatenated according to the text, and the text model is made. During the training phase, each speaker utters about ten sentences, and the sequences for feature parameters (cepstrum and so on) are extracted. Then, the HMM (for example, a 3-state, 256-mixture tied-mixture HMM) for each phoneme (a short utterance unit such as /a/, /i/) is made.

In this method, how to create speaker-specific phoneme models with accurate speaker and phonetic information is crucially important. About ten sentences are inadequate as training data in order to estimate parameters of each phoneme model (HMM). In speaker verification, since a method that needs a large amount of training data is unrealistic, it is not practical to use a larger number of sentences. Therefore, speech data for multiple speakers (9,000 sentences uttered by 30 male and 30 female speakers) is used for creating speaker-independent phoneme models as initial mod-

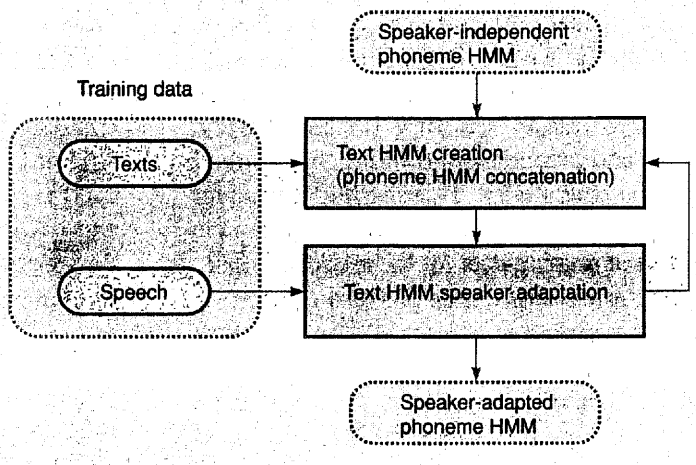


Figure 5. Method for Making Speaker-Adapted Phoneme Models

els. For each speaker, training data is applied to the initial models, and the model parameters are adapted to the speaker to create the speaker-specific phoneme models.

The procedure for making speaker-specific phoneme models is shown in Fig. 5. Since the texts of the training data are known, speaker-independent phoneme models are concatenated according to each text, and the text model is created. The speech data is applied to the model, and the model parameters (the mixture means and weighting factors) are estimated by using the Baum-Welch algorithm. Since it is difficult to estimate the mixture covariance matrices using a small amount of training data, the mixture covariance matrices of speaker-independent models are used for those of speaker-specific models. Then, the text model is divided into models of each phoneme. The above procedure is repeated several times for all training data and the phoneme model parameters are adapted to the speaker.

We have reported phoneme dependent/independent iterative speaker adaptation as a method that adapts speaker-independent phoneme models to the speaker. In this method, tied-mixtures are used as a common component, and two different types of training, phoneme-dependent training (as mentioned above) and independent training, are applied to them sequentially. This method can adapt phoneme models to the speaker even when the amount of training data for the model is small.

During the verification phase, the phoneme models for the claimed speaker are concatenated according to the prompted text, and the text model is made. The likelihood value for the input speech to match the text model is calculated and compared with the threshold for speaker decision. When using 15 speakers' voices recorded over 10 months, the above adaptation method achieved a speaker and text verification rate of 99.3%.

## 2 Technical Issues

### *2.1 Overview*

In speaker verification, there remain a lot of technical problems that should be investigated. For example, one big problem for speaker verification systems is that a person's voice changes with time. Some mechanisms that regularly update a reference pattern or model for each speaker should be considered. Moreover, there are few studies that investigate a method of setting a proper threshold for speaker verification beforehand. Few studies have reported results of performances for speaker verification in real conditions such as with noise, many speakers, and speech on telephone lines. The following sections introduce technical issues: the combination of multiple speech features, calculation of reliable average distance or distortion, and likelihood normalization related to our approaches.

### *2.2 Combination of Multiple Speech Features*

Two types of feature parameters can be extracted from voiced speech: one is for the spectral envelope, which represents the resonance and antiresonance characteristics of the vocal tract, and the other is for the spectral fine structure, which represents the voice source information of the vocal cords. Cepstrum and delta-cepstrum (the first-order regression coefficient of cepstrum) are used as typical parameters of the former and pitch (fundamental frequency) and delta-pitch (the first-order regression coefficient of pitch) are used as typical parameters of the latter. As shown in Section 2, the feature parameters for the spectral envelope, especially cepstrum, are used in speaker verification. This is because, as many experiments have shown, high performance can be obtained by using cepstrum, and it can be extracted stably. On the other hand, pitch can only be extracted from voiced parts of speech; it cannot be extracted for unvoiced parts. Moreover, it is difficult to extract pitch even from voiced parts accurately and automatically. Although the difference in pitch between speakers is large, it is also difficult to use pitch efficiently because a person can greatly vary the pitch of his/her voice.

On the other hand, a person uses several features, including tone, pitch, and loudness, to identify voice characteristics and can cope

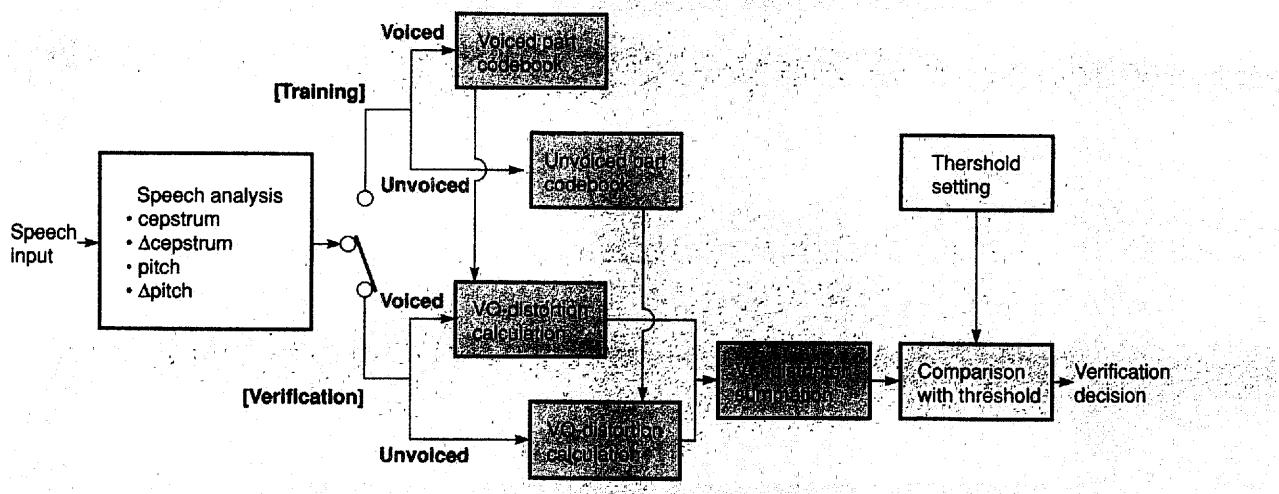


Figure 6. Text-Independent Speaker Recognition Based on VQ Using Vocal Tract and Source Information

with text-dependent and temporal spectral variations. This suggests that a combination of speech features will improve speaker verification.

We previously proposed a method that uses both the cepstrum and pitch parameters. Pitch is extracted only from the reliable voiced parts. Fig. 6 shows a system block diagram of this text-independent speaker verification method based on VQ. Two codebooks are made for each speaker: a voiced codebook and an unvoiced codebook. A voiced-code vector consists of cepstral and delta-cepstral coefficients, and pitch and delta-pitch frequencies, whereas an unvoiced-code vector consists only of cepstral and delta-cepstral coefficients. Input speech is classified into voiced/unvoiced parts by using a voiced/unvoiced decision and vector-quantized using the voiced/unvoiced codebooks. The VQ-distortion values are accumulated over all frames assigned to each codebook. The values for both the voiced and unvoiced codebooks are summed and used for speaker decision. This method, as evaluated using a nine-speaker database recorded over three years, achieved 98.7% speaker verification accuracy.

In general, speech features can be divided into static features (cepstrum, pitch, and so on) and dynamic features (delta-cepstrum, delta-pitch, and so on). Several papers have shown that delta-cepstrum can be efficiently used with cepstrum for text-dependent verification. However we know of no papers showing that the combination of cepstrum and delta-cepstrum improves text-independent verification based on HMM. Further study of text-independent verification is needed.

### *2.3 Calculation of Reliable Average Distance or Distortion*

The distribution of feature vectors extracted from speech varies with the texts. In text-independent verification, since an arbitrary text is uttered, the text can include phonemes that are not included in the training text. Feature vectors that correspond to the input speech can be far from the distribution of the training data even if the speech is uttered by the true speaker. They thus become noise in calculating the average VQ-distortion between the input speech and the claimed speaker's codebook or in calculating the likelihood value for the claimed speaker's HMM of the input speech. Moreover, physical characteristics of speech vary from session to session, even for the same

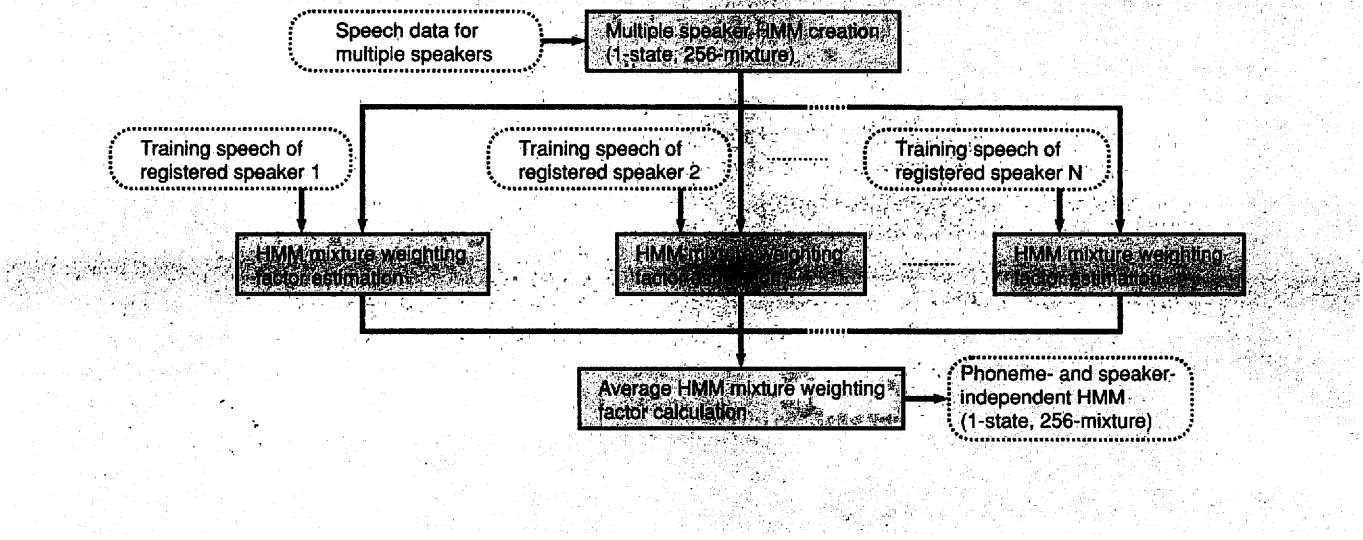


Figure 7. Method for Making Phoneme- and Speaker-Independent HMM

text and speaker.

Our proposed distance measure is characterized by selective matching using only a stable subset of the test speech in calculating the VQ-distortion. The error rates in speaker verification using this method, as evaluated using our nine-speaker database recorded over three years, are roughly 50% of those using all speech parts.

This approach can be applied to not only text-independent speaker verification, but also to a wide range of pattern verification problems that include unpredictable variations.

### *2.4 Likelihood Normalization*

The VQ-distortion/likelihood value for input speech to match the claimed speaker's codebook/HMM has a wide range for different texts spoken at different times, even by the same speaker, so it is difficult to set stable thresholds. In speaker verification based on HMM, we have reported a normalization method based on a posteriori probability, which is given by

$$P(s_c|x) = \frac{P(x|s_c) \times P(s_c)}{\sum_i \{P(x|s_i) \times P(s_i)\}} \cong \frac{P(x|s_c)}{\sum_i P(x|s_i)}$$

where $s_i$ is a speaker and $s_c$ is the claimed speaker. The $P(s_i)$ is the probability for speaker $i$, and is assumed to be a constant for all speakers. The $P(x|s_c)$ is the likelihood value of a feature vector $x$ given claimed speaker's HMM. The amount of calculation for $\sum P(x|s_i)$ is enormous since the likelihood values between the input speech and the models for a large number of speakers must be calculated. To cope with this problem, we proposed a method that approximates this summation by a single likelihood value obtained using a phoneme- and speaker-independent model made by using all registered speakers and texts.

The method for making the phoneme- and speaker-independent model is as follows (Fig. 7). Speech data for multiple speakers are used to create an initial model (a 1-state, 256-mixture Gaussian HMM) that universally covers the feature space of speakers. We used 9,000 sentences uttered by 30 male and 30 female speakers who are different from the registered speakers, as the speech data. In order to adapt the initial model to the overall characteristics of all the registered speakers, the following procedures are performed. For each registered speaker, training speech is applied to the

multiple-speaker HMM and the mixture-weighting factors are estimated by using the Baum-Welch algorithm. This procedure is repeated for all registered speakers. Then, for each mixture, the weighting factors calculated for all of the registered speakers are averaged to create a phoneme- and speaker-independent HMM (1-state, 256-mixture HMM). When a new speaker is added as a registered speaker, the phoneme- and speaker-independent HMM is updated by estimating the mixture-weighting factors of the multiple-speaker HMM using the training speech of the new speaker and recalculating the average of the weighting factors for all speakers including the new speaker. This updating procedure needs only a small amount of computation. Using 15 speakers, the speaker verification error rate was about half of that achieved without using this likelihood normalization method.

### *2.5 Database*

Speaker verification technology has been studied continuously by AT&T, TI, ITT, and BBN in the U.S., Swansea Univ. in the U.K., Telecom-Paris in France, and NTT in Japan. Until now, they have all used their own databases and separately evaluated their methods. In order to compare the performance of these different methods, a common database is needed. Since the amount of work involved in making a big database is enormous, a project for making a common database is now under consideration by many of these organizations.

## 3 Applications

### *3.1 Text-Dependent Method*

In the U.S., telephones that accept credit cards are popular, and there is a serious problem with stolen or misused cards. US Sprint cooperated with Texas Instruments (TI), and using TI's speaker verification technology, added a speaker verification function to the credit card service (Voice Phone Card) in order to improve its security. This service was started in January 1994. Although the performance of the speaker verification function itself is not so high, the total probability of a user losing his/her card, the lost card being misused, and an impostor being accepted is very small, so the security is very high. In practice, a user utters his/her ten digit social security number (every citizen of the U.S. has one), in response to the instructions from the system. After he/she has been identified, he/she utters for example "call Miyawaki (the name "Miyawaki" has been registered beforehand)," and the system make a telephone call to Miyawaki. Since the text for speaker verification is fixed for each user in this method, it is a text-dependent method.

### *3.2 Hybrid of Text-Dependent and Text-Prompted Methods*

AT&T has performed field tests in order to add a speaker verification function to general-purpose credit cards and telephone-specific credit cards since 1993. A user utters 10 digits or 7 short words, and each voice pattern for each utterance is stored in a memory IC in the card. The speaker verification system has a function for matching input speech and the voice patterns in the card. During the verification phase, the system selects one word randomly and prompts it to the user. The voice is accepted only when the true user utters the word correctly. In this method, although the set of words for each user is fixed, since one word is selected randomly and prompted at every verification, this method is intermediate between text-dependent and text-prompted methods.

## 4 Conclusions

This paper reviewed basic methods of speaker verification, especially text-dependent, text-independent, and text-prompted verification, several technical problems, and some applications. When using cepstral coefficients extracted from speech data recorded in a quiet room with a high-quality microphone, the text-independent speaker verification rate is currently 99.8% and text-prompted speaker verification rate is 98.9%. For the next step, which is practical application, speaker verification methods should be evaluated by using

speech data on real telephone lines that include noise. Moreover, although the verification performance is theoretically not related to the number of registered speakers, since a binary decision as to whether or not the input speech is uttered by the claimed speaker is done in speaker verification, in practice, experiments using a large number of the registered speakers should be examined.

*References*


(1) S. Furui: "An Overview of Speaker Recognition Technology," Proc. ESCA Workshop on Automatic Speaker Recognition Identification Verification, pp.1-9, 1994.

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

(3) S. Furui: "Cepstral Analysis Technique for Automatic Speaker Verification," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-29,2, pp.254-272, 1981.

(4) J.M. Naik, L.P. Netsch and G.R. Doddington: "Speaker Verification over Long Distance Telephone Lines," Proc. IEEE Int. Conf.Acoust., Speech, Signal Processing, S 10 b. 3, Glasgow, 1989.

(5) T. Matsui and S. Furui: "Comparison of Text-Independent Speaker Recognition Methods Using VQ-Distortion and Discrete/Continuous HMM's," IEEE Trans. Acoust., Speech, Signal Processing, Vol. 2, No. 3, pp.456-459,1994.

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

(7) F.K. Soong and A.E. Rosenberg: "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition," Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, Tokyo, Japan, 17.5, pp.877-880, 1986.

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

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

(10) T. Matsui and S. Furui: "Speaker Adaptation of Tied-Mixture-Based Phoneme Models for Text-Prompted Speaker Recognition," Proc. IEEE Int. Conf.Acoust., Speech, Signal Processing, Adelaide, pp.I-125-128, 1994.

(11) T. Matsui and S. Furui: "A Text-Independent Speaker Recognition Method Robust against Utterance Variations," Proc. IEEE Int. Conf.Acoust., Speech, Signal Processing, pp.281-284, 1991.

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


## The Authors


**Tomoko Matsui**
Research Scientist, Furui Research Laboratory, Human Interface Laboratories, NTT, currently engaged in research on speaker and speech recognition technology.



**Sadaoki Furui**
Research Fellow and the Director, Furui Research Laboratory, Human Interface Laboratories, NTT.