Chapter 5

PERSONALIZE MOBILE ACCESS BY SPEAKER AUTHENTICATION

Ke Chen
School of Computer Science
The University of Birmingham
Edgbaston, Birmingham B15 2TT
United Kingdom
K.Chen@cs.bham.ac.uk

Abstract In recent years wireless networks have been rapidly grown up, which leads to the possibility of pervasive access to information systems. At present, the most commonly available ubiquitous access device to the network is still mobile telephone. In particular, cellular phone can be used for ubiquitous access anytime and anywhere and, therefore, the only ubiquitous user access mode is spoken language. Obviously, both a cellular phone handset and any private information access or electronic transaction demand to be protected from being stolen or broken in, which paves the way for personalized services. In this chapter, we envisage a bilateral user authentication framework for applications in wireless environments. On the one hand, we use text-dependent speaker verification for handset protection as the primary stage of our security system. On the other hand, a more sophisticated speaker authentication system consisting of text-independent speaker verification and verbal information verification is located in the authentication center of a server site for further protection. Our framework attempts to derive maximum synergy from biometric and non-biometric speech technologies without loss of easy-to-access properties. Under this framework, we have conducted some experiments by enabling component technologies in terms of Mandarin Chinese. Our simulation results indicate that the enabling component techniques to support this framework are ready to build such an authentication system for applications to personalized mobile access.

Keywords: Speaker authentication, personalized mobile access, speaker verification, text-dependent, text-independent, verbal information verification, bilateral authentication, cellular phone, speech information system, server-client architecture

5.1. Introduction

In recent years wireless networks have been rapidly developed and become indispensable components in the telecommunication world. According to the literature [11], there are over 100 million cellular and personal communication service phones in use as of mid-2000 in the United States. Even in China, a developing country, current estimates of China Telecom Inc. indicate that there are over 70 million registered cellular phone users at the mid of 2001 and, moreover, such a market will be grown up rapidly. Thus, fraud becomes a common yet serious problem that the cellular operating companies have to tackle. In this circumstance, a fraud user is able to steal handsets or the phone and serial numbers from the airwaves. Then they could use these numbers in cellular phones and make long distance and even international calls. As a result, the cellular providers may have to pay unexpected costs of millions of dollars on a monthly basis [4].

Although the world of telecommunication in the future will be the seamless integration of real-time multimodal communications in a single network, the most commonly available ubiquitous access device to the network is cellular phone for ubiquitous access anytime and anywhere. Therefore, the only ubiquitous user access mode is spoken language, a natural mechanism for information access. Given more and more services such as mobile stock quotes and transactions are popular, security upon access becomes an unavoidable problem for a speech information system of a private or confidential nature. How to authenticate a user becomes critical to prevent unauthorized users from mobile access to the private information conveyed in speech.

To solve the security problems existing in wireless environments, there are two common approaches by the use of fraud *personal identification number* (PIN) features and mathematical authentication technologies. A PIN feature is usually a string consisting of digits or alphabets, which uniquely identifies a specific person. In this way, therefore, different PINs are assigned to those authorized or registered users. When such a user would like to make an outgoing call or access a speech information system, he/she has to first unlock his/her cellular phone or pass an authentication processing prior to any access by using the PIN. On the other hand, mathematical algorithms provide a powerful tool for protecting cellular phones. In this way, the phone is identified not only by the phone number and the serial number but also by a a random key. This random key is loaded in the cellular phone by the vendor. Once a new user is registered in the network, the same random key is loaded in the authentication center. When the user makes the first call, the cellular phone performs some calculations, by certain algorithms, in terms of the phone number, the serial number, and the random key. The results generated are transmitted in the airwaves. Accordingly, the authentication center takes the same operations by using the same inputs and algorithms. As a consequence, user authentication is done by checking the consistency between bilateral results; that is, the user is allowed to make the call only if two results are identical. Although the aforementioned approaches can reduce fraud activity, there exist explicit weaknesses. The use of PIN features seems troublesome and unnatural; a user has to wait until the cellular phone is unlocked to make a call. Moreover, clones are still able to steal the user's information from the airwaves. As a result, the clones eventually will be able to crack the codes even if such algorithms may be complicated. When the handset is occupied by an unauthorized person who knows the right PIN, the immediate loss seems unavoidable.

Recently, systematic studies have shown that *biometrics* provides an alternative yet natural way for user authentication [16, 32]. Biometrics handles authentication of individuals on the basis of biological and/or behavioral characteristics. In contrast to the traditional authentication approaches, the primary advantage is that biometrics cannot be misplaced and forgotten since biometric features are always inherently associated with human beings. As summarized in the literature [16], biometrics has a number of salient and desirable properties as follows: a) universality, b) uniqueness, c) permanence, d) collectability, e) performance, f) acceptability, and g) circumvention. There are numerous biometric features used for authentication. However, each of them is of its strengths and limitations in terms of the above properties and has to appeal to a special authentication application. In our circumstance, voice print or speech becomes the biometric feature available only. According to perception of biometrics experts [16], voice print is of the following properties: a) medium universality, b) low uniqueness, c) low permanence, d) medium collectability, e) low performance, f) high acceptability, and g) low circumvention. The properties of voice print provide a two-fold insight. On the one hand, high acceptability suggests that voice print be natural and become the ideal feature for user authentication in wireless environments. On the other hand, other unsatisfactory properties indicate that the voice print itself is insufficient to be a unique feature to perform user authentication in the environments in question. Thus, it poses a dilemma to us, and a solution to this dilemma is demanded such that user authentication can be performed by using only speech without loss of its desirable advantage.

As one of automatic biometric technologies, automatic speaker recognition has been studied for several decades [15, 21]. In general, speaker recognition is classified into two categories: speaker identification, a process of identifying an unknown voice token as belonging to one of registered speakers, and speaker verification, a process of accepting or rejecting the identity claim of a speaker. Apparently, speaker verification is more appropriate to user authentication in most circumstances. Moreover, a speaker verification system often works in either of two operating modes: text-dependent and text-independent. By text-dependent, the same or known text is used for training and test. In contrast, any text is allowed to be uttered in the process of either training or test in the text-independent mode. By comparison, a text-dependent system is conceptually simple yet inflexible while a text-independent system seems complicated yet flexible. Moreover, the performance of a text-dependent system is often reasonably better than that of a text-independent system while the text-independent system can perform in a more secure way if the user is allowed to speak any random phrase. No matter what the operating mode is, speaker recognition theoretically belongs to non-verbal speech classification since the information of speaker's characteristics conveyed in speech waves plays a crucial role in this process rather than those verbal contents carried by speech waves. As a consequence, speaker verification provides a reasonably good measure of security for access to a wireless network and to private/confidential information during a personalized service.

As a matter of fact, the voice print is inherently subject to change and sensitive to environments, which leads to a classification task of miscellaneous mismatches. Thus, the use of voice print itself fails to yield the desirable performance in contrast to other biometric features. Nevertheless, speech recognition, a verbal-content based speech classification task, has been well studied and received satisfactory performance [17], which makes an automatic telephone-banking like process feasible. Borrowing the telephone-banking concept, Li *et al.* first propose an alternative speech-based authentication approach – *verbal information verification* [22]. Other than the traditional speaker recognition, verbal information verification is a process that verifies spoken utterances against the pre-registered information in a personal profile. For user authentication, a verbal information verification system mainly inspects the verbal content conveyed in speech signals while a speaker recognition system takes advantage of a speaker's characteristics represented by the speech feature vectors [23]. Although verbal information verification has nothing to do with biometrics, it has generated considerably better performance (even error free) in user authentication [23, 24] assuming that the personal information is not stolen by unauthorized people. Therefore, the combination of traditional speaker verification and verbal information verification provides a promising way for high-performance user authentication without loss of the desirable property, easy-to-access, of speech.

In this chapter, we propose a bilateral user authentication framework though the combination of speaker verification and verbal information verification for personalized mobile access to private/authorized device, e.g. cellular phone, and confidential information, e.g. electronic financial transaction. In wireless environments, client device and distributed authentication centers constitute a server-client network and, dependent upon different tasks, user authentication is performed in either of two sites or both. Considering complexity in implementation and acceptability, text-dependent speaker verification is used in client device for primary user authentication, while the combination of text-independent speaker verification and verbal information verification leads to an innovative user authentication procedure in authentication centers. Thus, the hierarchical and interactive authentication schemes constitute a new user authentication framework for personalized mobile access in wireless environments. Such a framework could provide a potential solution to the aforementioned dilemma towards an error-reduction and easy-to-access service in personalized mobile access. In terms of Mandarin Chinese dialect, we have investigated the enabling component technologies. Our experimental results indicate that the major component technologies to support this framework are ready for real use though there are challenging implementation issues to be studied in the future.

The remainder of this chapter is organized as follows. Section 5.2 presents the bilateral user authentication framework. Section 5.3 describes key enabling component technologies developed in terms of Mandarin Chinese dialect. Section 5.4 reports experimental results, and the last section draws conclusions.

5.2. Bilateral User Authentication Framework

In this section, we present a bilateral user authentication framework to personalize mobile access. On the basis of the framework, moreover, we describe a scenario example for personalizing mobile access.

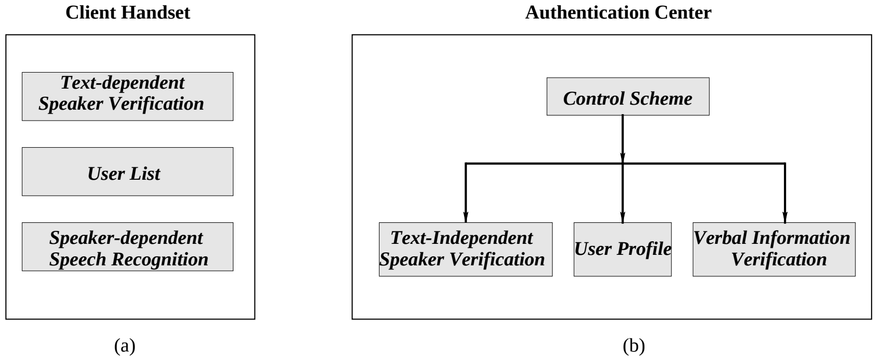

Client Handset
Text-dependent Speaker Verification
User List
Speaker-dependent Speech Recognition
Authentication Center
Control Scheme
Text-Independent Speaker Verification
User Profile
Verbal Information Verification
(a)
(b)

Figure 5.1. The schematic diagram of a bilateral user authentication. (a) The client handset. (b) The authentication center.

As illustrated in Figure 5.1, the bilateral user authentication framework consists of two modules: *client device* and *authentication center*. The client device module is used for primary user authentication in order to enable the cellular phone to work for authorized users. The authentication center module works for further authentication when a user attempts to place a call of expensive cost or access to an information system containing private or personal information.

The communication between two modules is through the use of a simple yet special protocol such that the components in the authentication center can be activated.

In the client device module, there are three components related to user authentication as shown in Figure 5.1(a). A text-dependent speaker verification scheme is the key component to determine whether a user is given the right to access to the cellular phone. It should be pointed out that the claim process can be neglected since a cellular phone is usually assumed to belong to a specific user. Thus, the serial number of the phone provides a way to automatically claim identity. In case that a cellular phone may be shared by a small group of people, e.g. people belonging to a family, the user list registers those authorized users who are allowed to use this cellular phone handset. It should be stated that how to organize a user list well is a research topic and some potential solutions have been raised [6]. The speaker-dependent speech recognition scheme provides a set of simple speech recognition engines for voice-based dialing according to an address book. Note that due to the accessibility of multiple users there are an independent speaker verification/speech recognition schemes for the users enrolled in the user list. Thus, the population in the user list should not be large.

The authentication center provides a strict and final authentication mechanism for each user prior to some important mobile access. There are four schemes in the module for such authentication as depicted in Figure 5.1(b). The control scheme is used to globally control all the authentication mechanism and selectively activate an authentication scheme. Once it is activated, the text-independent speaker verification scheme always probes the specific user's identity during conversation to see if the identity of the current user occupying the cellular phone is consistent with that of its registered users claimed automatically. The user profile scheme stores the files of all the registered users served by the authentication. The contents of each file includes the personal and private information of each registered users, which provides the basis for verbal information verification. When a user is initially registered as a new user in the authentication center or the speaker verification scheme reports an inconsistent result, the verbal information verification scheme will be invoked. As a result, the current user is asked to answer a set of questions randomly selected from an elaborate questionnaire. Only if all the answers given are correct, the user will be allowed to make a continuous access. For security, the user profile could be updated regularly.

In order to intuitively understand our framework, we give a scenario example of personalized mobile access, which demonstrates how our framework works. For a new user, enrollment in the client device and the authentication center becomes the first step. In the client device site, the user is asked to utter voice-based commands and names to be dialed three times. In the authentication center, the enrollment process is to finish a user profile, through filling out a form, including personal and private information and a set of self-defined answers to some questions.

Once the enrollment has been done, the user may start the personalized access to his/her cellular phone. After the power is turned on, the user list scheme is activated, and thus, a list of authorized users are shown on the panel of the cellular phone where each authorized user is labeled by a number. The user utters the number represented him/her for identity claim. A special case is that the cellular phone is owned by one authorized user. In this circumstance, the identity claim is default and the user list is not shown. At this moment, the cellular phone is still locked. Prior to access to the phone, the user has to unlock the phone by a voice-based command. When the utterance of this command is achieved, the text-dependent speaker verification scheme is activated and authenticates the current user based on the templates stored in the enrollment process. If the user's identity is authenticated, the cellular phone is ready to enter any conversation phase. In order to be easy-to-access, a user is encouraged to make a phone call by using the voice-based dialing. Thus, the speaker-dependent speech recognition engine is activated for this task. After a conversation is performed, the phone may be locked again by the user through use of the voice-based command. There are the following cases for the client device to send a request to the authentication center: (1) the first time a user takes the cellular phone, (2) failure to unlock the phone or to dial by voice after three trails, and (3) making a long distance phone call or access to the private or personal information.

As illustrated in Figure 5.1(b), there are two authentication schemes in the authentication center. During the enrollment, users provide the corresponding user profiles such that the verbal information verification scheme can work for any registered users. When a user makes his/her first phone call, the verbal information verification scheme is activated by an automatic request from the client device. Once the user identity is verified by the verbal information verification scheme, a text-independent speaker verification model is created for this user. During the first conversation, all the utterances are automatically used to train the speaker model. That is, the second authentication scheme, text-independent speaker verification, is created based on the verbal information verification scheme during the first phone call. Once the text-independent authentication scheme is created, it would be activated by any long distance phone call request from the client device site. The text-independent authentication scheme inspects the phone call by report a verification result in a fixed interval. If the verification result indicates that an impostor is accessing the personalized client device, the phone call is immediately suspended. It is followed by a verbal information verification test. If the test is successful, the phone call is activated again. As a consequence, the utterances during the verbal information verification and conversations thereafter are used to update the text-independent speaker model. Note that in order to ensure a low error rate the text-independent speaker model is always updated in an autonomous way if the user's identity is verified. Thus, a personalized mobile access is carried out by the bilateral user authentication framework.

In the sequel, we are going to present some enabling technologies to support our bilateral user authentication framework for personalized mobile access.

5.3. Enabling Speaker Authentication Technologies

In this section, we present enabling speaker authentication technologies to support our bilateral user authentication framework. We first describe the speaker verification technologies used in the client device and the authentication center. Then, we present a verbal information verification technology in terms of Mandarin Chinese dialect. Finally, we discuss how to derive maximum synergy for user authentication from both text-independent speaker verification, a biometric technology, and verbal information verification, a non-biometric technology.

5.3.1 Speaker Verification

Speaker verification is a biometric authentication technology. As illustrated in Figure 5.2, the critical technical components in speaker verification include speaker modeling and decision-making strategies. There are numerous approaches to speaker verification [3, 29, 13, 25, 15, 14, 5]. For the use in our framework, the speaker modeling in text-dependent speaker recognition tends to be as simple as possible and computationally efficient in the client device site, while speaker modeling in text-independent speaker recognition would be demanded to produce the error rate as low as possible. In addition, decision-making strategies used are different in the client device and the authentication center. Here, we present two enabling technologies to meet our requirements.

NFL-based Text-Dependent Speaker Verification. Theoretically, speaker verification belongs to the category of non-verbal speech classification regardless of operating modes. However, most of text-dependent speaker verification approaches take advantage of verbal contents to capture speaker's characteristics [15], which often needs the strict temporal alignment. A temporal alignment process usually suffers from a high computational load, e.g. dynamic time warping [30]. Our previous studies showed that some instantaneous information carried by certain frames within an utterance can play a more important role in text-dependent speaker recognition and the use of transitional (inter-frame) information may not be involved in a strict temporal alignment [10]. Our recent studies indicated that the *nearest feature line* (NFL) is able to be a

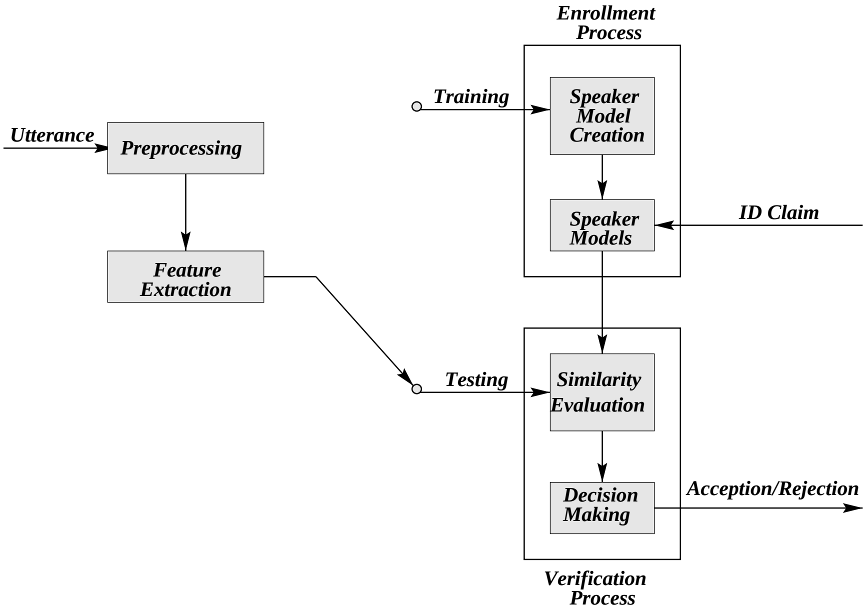


Figure 5.2. The schematic diagram of a typical speaker verification system.

text-dependent speaker verification technique without the strict temporal alignment, which does not involve in high computational load and, in particular, the performance of an NFL-based text-dependent system is better than that of a dynamic temporal warping system [9]. Thus, the NFL-based text-dependent speaker verification approach qualifies as the enabling technology in our framework.

The NFL assumes that there are at least two prototypes (in our case, two utterances of a fixed phrase) for each speaker. The line passing through two feature points, extracting from two utterances after preprocessing, can extrapolate or interpolate to form a line, named by *feature line* in the NFL approach, in the feature space. Now we consider two feature points, $\mathbf{x}_i^s$ and $\mathbf{x}_j^s$, belonging to speaker s. Thus, the distance d between the feature line $\overline{\mathbf{x}_i^s\mathbf{x}_j^s}$ passing through $\mathbf{x}_i^s$ and $\mathbf{x}_j^s$ and a query point $\mathbf{x}_q$ is calculated by

$$d(\mathbf{x}_q, \overline{\mathbf{x}_i^s\mathbf{x}_j^s}) = ||\mathbf{x}_q - \mathbf{p}_{i,j}^s||. \tag{5.1}$$

Here $\mathbf{p}_{i,j}^s$ is a point on the feature line achieved by projecting $\mathbf{x}_q$ to $\overline{\mathbf{x}_i^s\mathbf{x}_j^s}$. As a result, such a point can be obtained by linearly combining two feature points in terms of the query point as follows:

$$\mathbf{p}_{i,j}^s = \mu\mathbf{x}_i^s + (1-\mu)\mathbf{x}_j^s, \tag{5.2}$$

where

$$\mu = \frac{(\mathbf{x}_q - \mathbf{x}_i^s)^T(\mathbf{x}_j^s - \mathbf{x}_i^s)}{(\mathbf{x}_j^s - \mathbf{x}_i^s)^T(\mathbf{x}_j^s - \mathbf{x}_i^s)}.$$

For speaker s, any pair of his/her feature points constitute a feature line. For a given query point, $\mathbf{x}_q$, there is the nearest feature line, $\overline{\mathbf{x}_{i^*}^s \mathbf{x}_{j^*}^s}$, achieved by

$$\overline{\mathbf{x}_{i^*}^s \mathbf{x}_{j^*}^s} = \arg\min_{i,j} d(\mathbf{x}_q, \overline{\mathbf{x}_i^s \mathbf{x}_j^s}). \tag{5.3}$$

To build an NFL speaker model, we need to extract feature points or prototypes from raw speech data. In our method, each utterance, corresponding to a fixed phrase, is modeled to form a prototype (for details, see Section 5.4.2). Thus, the NFL speaker model is created by constructing the feature line space through the combination of prototypes in a pair-by-pair way. Note that the aforementioned prototypes should be normalized prior to forming the feature line space in order to facilitate the decision-making described later on.

Once a speaker model is built, the remaining task is how to make a right decision for an unknown voice token. In our circumstance, there are only enrollment data from the registered user himself/herself and no other speech data available since we allow a user to flexibly choose any phrase as the text. Obviously, we cannot use any traditional method to set a threshold and build either a background or a cohort model [15]. Cohort modeling is a typical approach to train a background model for decision-making in speaker verification. The idea underlying this approach is to build a model by the use of speakers who have acoustic characteristics similar to a specific speaker. Once a cohort model is available, the decision-making can be performed by comparing scores produced by the speaker model with that by his/her corresponding cohort model. Although a cohort model is merely available by associating with other speakers, recent studies demonstrated that the use of only enrollment data to build a background model, hereinafter named by *pseudo-cohort model*, leads to the appreciably good performance [31]. Motivated by this work, we build such pseudo-cohort models in terms of an NFL speaker model for each speaker by perturbing statistical components in his/her speaker model (for details, see Section 5.4.2). As a consequence, in our system, the decision-making on the client handset site is performed by means of the speaker and the pseudo-cohort models.

GMM-based Text-Independent Speaker Verification. As a typical approach, Gaussian Mixture Model (GMM) has been used especially for text-independent speaker recognition to characterize speaker's voice in the form of probabilistic model. It has been reported that the GMM approach outperforms other classical methods for text-independent speaker recognition [28, 8]. Here, we briefly review the GMM-based speaker identification scheme that will be used, as a technical component in our authentication center.

For a feature vector denoted as $\mathbf{x}_t$ belonging to a specific speaker s, the GMM is a linear combination of K Gaussian components as follows:

$$P(\mathbf{x}_t|\lambda_s) = \sum_{k=1}^{K} \omega_{s,k} \; P(\mathbf{x}_t|\mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k}). \tag{5.4}$$

Here $\omega_{s,k}$ is a linear combination coefficient for speaker s $(s = 1, 2, ..., S)$. $P(\mathbf{x}_t|\mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k})$ is a Gaussian component parameterized by a mean vector, $\mathbf{m}_{s,k}$, and covariance matrix, $\mathbf{\Sigma}_{s,k}$ as follows:

$$P(\mathbf{x}_t|\mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{\Sigma}_{s,k}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}_t - \mathbf{m}_{s,k})^T \mathbf{\Sigma}_{s,k}^{-1} (\mathbf{x}_t - \mathbf{m}_{s,k})\right]. \tag{5.5}$$

Usually, a diagonal covariance matrix is used in Eq. (5.5). Given a sequence of feature vectors, $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_t, \cdots\}$, from a specific speaker's utterances, parameters estimation for $\lambda_s = (\omega_{s,k}, \mathbf{m}_{s,k}, \mathbf{\Sigma}_{s,k})$ $(k = 1, \cdots, K, s = 1, \cdots, S)$ is performed by the Expectation-Maximization (EM) algorithm. Thus, a specific speaker model is built through finding proper parameters in the GMM based on the speaker's own feature vectors.

To evaluate the performance, a sequence of feature vectors is divided into overlapping segments of T feature vectors for identification [28]:

$$\overbrace{\mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+T-1}}^{segment\ l}, \mathbf{x}_{l+T}, \cdots\cdots$$

$$\mathbf{x}_l, \overbrace{\mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+T-1}, \mathbf{x}_{l+T}}^{segment\ l+1}, \mathbf{x}_{l+T+1}, \cdots\cdots$$

For a testing segment $X^{(l)} = \{\mathbf{x}_l, \mathbf{x}_{l+1}, \cdots, \mathbf{x}_{l+T-1}\}$, the log-likelihood function of a GMM is as following:

$$\mathcal{L}(X^{(l)}, \lambda_s) = \sum_{t=l}^{l+T-1} \log P(\mathbf{x}_t|\lambda_s) \quad s = 1, \cdots, S. \tag{5.6}$$

Thus, the likelihood value, $\mathcal{L}(X^{(l)}, \lambda_s)$, is a score produced by the speaker model corresponding to the claimed identity which will be used for decision-making.

Unlike the client handset site, there may be a large amount of data belonging to other speakers available off-line, e.g. a standard speech corpus, in the authentication center site. For the purpose of decision-making, therefore, it is feasible to utilize the data for creating a speaker-independent background model. As a result, we adopt a GMM of numerous Gaussian components and a non-diagonal covariance matrix, $P(\mathbf{x_t}|\lambda_{SI})$, to form such a background model (for details, see Section 5.4.2). Similarly, the decision-making, corresponding to a testing speech segment, is performed by means of the GMM-based speaker and background models. For an utterance of several segments, the final decision-making is achieved by a majority voting on the basis of the decision-making results with respect to all the testing speech segments comprised of this utterance.

5.3.2 Verbal Information Verification

Verbal information verification is an authentication technology recently developed in speech processing community [21, 23] where a claimed speaker is accepted/rejected by verifying spoken utterances against the information stored in a given personal data profile. Strictly to say, this technology does not belong to biometrics because it uses only the contents carried in speech for authentication. As pointed out previously, there are a number of problems as speaker verification is applied in real world, e.g., acoustic mismatch, quality of the training data, inconvenience of enrollment, and the creation of a large database to memorize all the registered speaker patterns. Obviously, the use of verbal information verification is able to enhance speaker authentication technologies. Although verbal information verification is regardless of speakers' acoustic characteristics, the technology is highly dependent on a dialect since the contents carried in speech need to be verified. Here we present the verbal information verification technology in terms of Mandarin Chinese.

General Description. Although verbal information verification has been successful in English language, it is still questioned that such a technology is effectively applicable to other languages. Mandarin Chinese is the most widely used language in the world since there are around 1.3 billion Chinese native speakers. Previous studies [19, 20] showed that Mandarin Chinese is different from English in numerous aspects. Some salient features in Mandarin Chinese are summarized as follows. First, every word of Chinese has only one syllable and consists of explicit semi-syllable configuration; INITIAL and FINAL. INITIAL is always a consonant, while FINAL could be one of single vowels, compound vowels, and vowels along with consonants. Next, all the Chinese syllables include FINAL, while INITIAL may not be contained in a Chinese syllable. Unlike phoneme in English, INITIAL and FINAL are basic acoustic unit in Mandarin Chinese instead. Thus, we need to use them for acoustic modeling, which results in a large difference from other languages in acoustic modeling. Finally, there are a few of words that are commonly used but make no contribution to verbal information verification, such as 'year', 'month', and 'day' in Mandarin Chinese as a question about birthday is raised, since these words are always present in the answer regardless of speakers. In addition, the same meaning can be represented by an alternative word; e.g., ‘day’ can be spoken in two different ways in Mandarin Chinese. Such information is hardly captured from users’ private data profile. All the aforementioned problems are worth studying, which causes the Mandarin verbal information verification to become a challenging task.

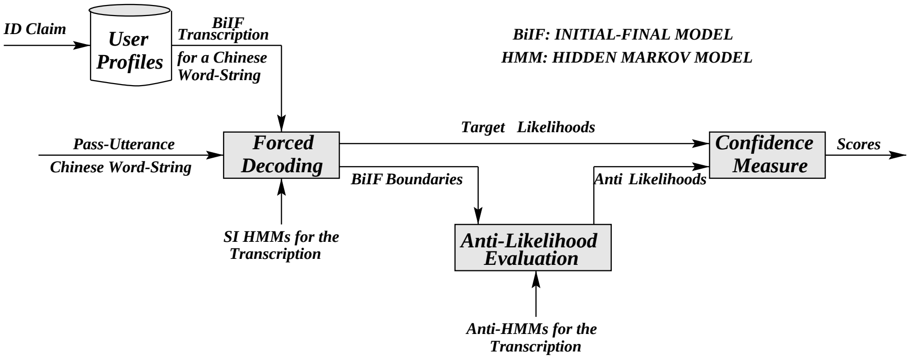


Figure 5.3. The schematic diagram of a Mandarin verbal information verification system.

For Mandarin verbal information verification, we have presented an architecture as depicted in Figure 5.3. Once an identity is claimed by a user, our system transcribes the pass-utterance from his/her private data profile. For instance, an answer to the question “What is your name?" is transcribed into a Chinese word string. During this transcription, the pass-utterance is acoustically modeled as a string consisting of INTIALs and FINALs. By the transcription, the system decodes the the pass-utterance. This process is summarized as *forced decoding* in Figure 5.3. As a result, forced decoding yields the INITIAL-FINAL’s segmentation boundaries for the string. Thus, the decision-making, accepting/rejecting the claimed speaker in terms of the utterance, can be performed by a hypothesis test.

According to Chinese linguistics, there are 22 INITIALs and 37 FINALs in Mandarin Chinese dialect. In order to model the co-articulation between them, we use the right context-dependent INITIAL-FINALs as basic acoustic unit, hereinafter abbreviated by *BiIF*. In our system, we employ hidden Markov models (HMMs) to model the acoustic units. As a result, an INITIAL-based BiIF model is a left-to-right (without jump) connected HMM of three states, while a FINAL-based BiIF model is a 5-state HMM of the same structure. By combination, totally, there are 1260 BiIF models in Mandarin Chinese. It is almost impossible to collect enough training data for those models. Instead we adopt a decision-tree based clustering method to reduce the number of states and models [24]. Moreover, we use segmental K-means and import decision tree algorithms together in on iterative step during training as done in the work [27]. In contrast, such a training method is more robust than the traditional decision-tree clustering algorithm. As a consequence, the speaker-independent acoustic models are achieved by fitting the training data to HMM models by the Viterbi learning algorithm [18, 24]. Similarly, we also use the same method to achieve anti-HMM models corresponding to those BiIF models by those data used for training all the target BiIF models of the same context. More details on implementation will be described in Section 5.4.3.

Decision-Making Procedure. Like speaker verification, decision-making is also involved into verbal information verification. In order to facilitate presentation, we first describe the verbal information verification process in a more formal way. Then, the decision-making procedure is presented based on the formal description.

Based on achieved HMMs and anti-HMMs, utterance segmentation is performed as follows. When our system prompts one single question at a moment, it knows the expected critical information, registered in his/her personal profile of the claimed speaker, to the prompted question and the corresponding subword sequence of N acoustic units, $\mathbf{S} = \{S_n\}_{n=1}^{N}$. Thus, the acoustic unit models, $\lambda_1, \cdots, \lambda_N$, in the same order of $\mathbf{S}$ are applied to decode the answer utterance in the forced decoding process. In this process, Viterbi algorithm is employed to find the maximum likelihood segmentation of the acoustic units, i.e.,

$$P(\mathbf{O}|\mathbf{S}) = \max_{t_1, t_2, \cdots, t_N} P(O_1^{t_1}|S_1) \cdots P(O_{t_1+1}^{t_2}|S_2) \cdots P(O_{t_{N-1}+1}^{t_N}|S_N), \quad (5.7)$$

where

$$\mathbf{O} = \{\mathbf{O}_1, \mathbf{O}_2, \cdots, \mathbf{O}_N\} = \left\{O_1^{t_1}, \cdots, O_{t_1+1}^{t_2}, \cdots, O_{t_{N-1}+1}^{t_N}\right\}. \quad (5.8)$$

Here $\mathbf{O}$ is a set of segmented feature vectors related to acoustic units, and $t_1, t_2, \cdots, t_N$ are the end frame numbers of acoustic unit segments. $\mathbf{O}_n = O_{t_{n-1}+1}^{t_n}$ is the segmented sequence of observations corresponding to the acoustic unit S_n from frame $t_{n-1}+1$ to frame t_n, where $t_1 \geq 1$ and $t_i > t_{i-1}$.

For a decoded acoustic unit, S_n, in an observed speech segment, $\mathbf{O}_n$, a decision-making strategy is demanded where the acoustic unit will be assigned to either hypotheses of H_0 and H_1. Here H_0 is the hypothesis that $\mathbf{O}_n$ is consistent with the corresponding items in the personal profile and H_1 is the alternative hypothesis. According to the Neyman-Person lemma [26, 12], the hypothesis test is described as

$$r(\mathbf{O}_n) = \frac{P(\mathbf{O}_n|H_0)}{P(\mathbf{O}_n|H_1)} = \frac{P(\mathbf{O}_n|\lambda_n)}{P(\mathbf{O}_n|\bar{\lambda}_n)}. \quad (5.9)$$

Here λ_n and $\bar{\lambda}_n$ are the target HMM and the corresponding anti-HMM for the acoustic unit, S_n. Thus, the *log-likelihood ratio* (LLR) for S_n is

$$R(\mathbf{O}_n) = \log r(\mathbf{O}_n) = \log P(\mathbf{O}_n|\lambda_n) - \log P(\mathbf{O}_n|\bar{\lambda}_n). \tag{5.10}$$

Accordingly, the averaging frame LLR, $\bar{R}_n$, is

$$\bar{R}_n = \frac{1}{L_n}\left[\log P(\mathbf{O}_n|\lambda_n) - \log P(\mathbf{O}_n|\bar{\lambda}_n)\right], \tag{5.11}$$

where L_n is the length of the speech segment. For each acoustic unit, a decision can be made by the following rule

$$\text{Acceptance}: \ \bar{R}_n \geq T_n; \quad \text{Rejection}: \ \bar{R}_n < T_n.$$

Here either an acoustic-unit dependent threshold, T_n, or a content-independent common threshold, T, can be determined numerically or experimentally.

Since a single utterance may contain numerous acoustic units, an utterance level decision is further needed to be made as well. For this purpose, we employ a normalized confidence measure as used in the work [23]. For an acoustic-unit string characterized by the INITIAL-FINAL model, λ_n, a confidence measure is defined as

$$C_n = \frac{\log P(\mathbf{O}_n|\lambda_n) - \log P(\mathbf{O}_n|\bar{\lambda}_n)}{\log P(\mathbf{O}_n|\bar{\lambda}_n)}, \tag{5.12}$$

where $P(\mathbf{O}_n|\bar{\lambda}_n) \neq 0$ indicates that this target score is larger than the anti-score and vice versa. Thus, a normalized confidence measure for an utterance of N acoustic units (subwords) as

$$\bar{C} = \frac{1}{N}\sum_{n=1}^{N} H(C_n) \tag{5.13}$$

Here, $H(C_n)$ is the Heaviside step function defined as

$$H(C_n) = \begin{cases} 1, & \text{if } C_n \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{5.14}$$

$\bar{C}$ is located in the fixed interval between zero and one. Due to the normalization in Eq. (5.12), the threshold, θ, is content-independent that can be determined separately. According to Eq. (5.14), an acoustic unit is accepted only if its C_n score is not less than the threshold θ. In other words, only those acoustic unit of $C_n \geq \theta$ can make a contribution to acceptance. As a result, $\bar{C}$ would be viewed as the percentage of acceptable acoustic units in an utterance. Hence, an utterance threshold may be set or adjusted in terms of the specification of our system and performance.

For verbal information verification, a test may include a number of utterances corresponding to the answers to several questions randomly selected from an elaborate questionnaire. Therefore, the sequential utterance verification must be considered for real applications. Fortunately, the above single utterance decision-making strategy, defined in Eqs. (5.12)-(5.14), can be directly extended to a sequence of subsets, which is similar to the step-down procedure in statistics [1]. Each of the subsets is an independent single utterance verification. As long as a subset is rejected, H_1 is chosen to be true and the testing procedure is terminated. In contrast, the claimed identity (user) is acceptable only if every subset passes the test, i.e., each H_0 is accepted.

5.3.3 Combination of Speaker and Verbal Information Verification

As presented in Section 5.2, the authentication center adopts a new speaker authentication strategy by combining text-independent speaker verification and verbal information verification. In such a strategy, the authentication center works in a natural way; the text-independent speaker verification scheme performs identity authentication in an automatic and transparency way, while the verbal information verification scheme is activated only if those circumstances listed in Section 5.2 occurs.

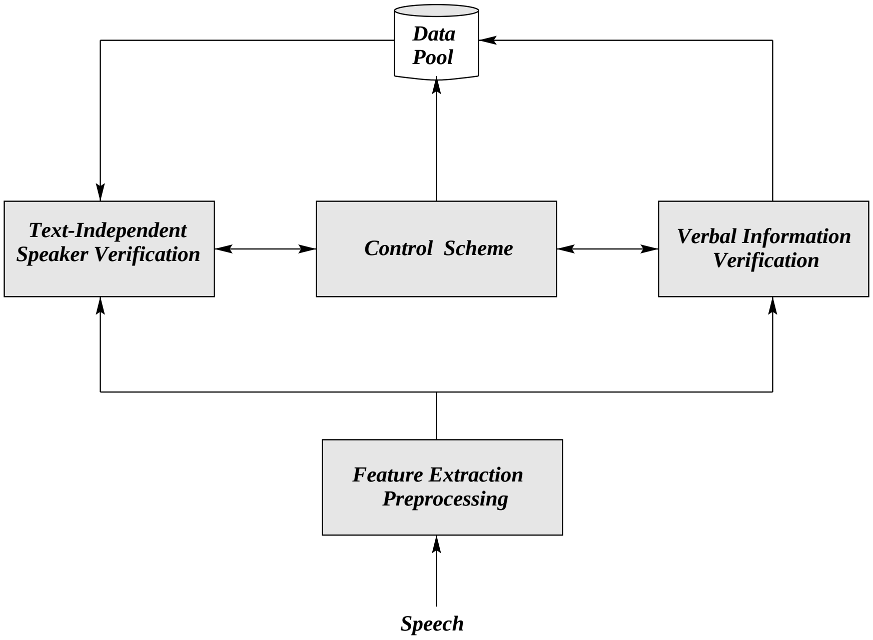

Data Pool
Text-Independent Speaker Verification
Control Scheme
Verbal Information Verification
Feature Extraction Preprocessing
Speech

Figure 5.4. The schematic diagram of a combination mechanism in the authentication center.

As illustrated in Figure 5.4, the control scheme is the key component of this combination scheme. When a special call request, e.g. international call, is received, the control scheme is activated and its default mode is to activate the text-independent speaker verification system based on the information of the registered caller. Thus, all the utterance in the current call is monitored by the text-independent speaker verification system during this phone call. Once a speech stream of a certain length is rejected, the system immediately breaks the call and keeps the transaction information, e.g. dialed number, in the meanwhile. Then, the control scheme suspends the text-independent speaker verification system and activates the verbal information verification system instead. Thus, a verbal information verification process proceeds; a number of questions are asked one by one and the answers from the current caller are verified based on the private profile of the handset owner. The answering utterances are stored in the data pool as shown in Figure 5.4. If any of the caller's answers is inconsistent with the corresponding item in the private profile twice, the control scheme also suspends the verbal verification system. Thus, the authentication center will terminate this call and put the memo information in the logout file. Otherwise, the control scheme redials the memorized number and makes a valid connection. In addition, the control scheme directs the speaker model in the text-independent speaker verification system to be adapted on the data available in the data pool. After the adaptation is performed, the control scheme will empty out the data pool.

It is worth mentioning that there are several differences between the combination strategy here and that proposed in the work [23]. First, two different verification systems in our combination strategy do not work simultaneously. Instead they work in an alternate way, while two verification systems in their approach have to work in a cascade way [23]. Next a text-independent speaker verification system can be used in our strategy, while a text-dependent speaker verification system is merely used for that combination [23]. Finally, automatic enrollment in the component speaker verification system is quite different. In their work [23], the speaker verification system can be trained only if the verbal verification system works for several times. In contrast, the data for automatic enrollment can be achieved while a valid user makes a local phone call. Furthermore, sufficient data can be collected from multiple calling phases such that the speaker model in the text-independent speaker verification system can be adapted regularly. Thus, the mismatch problem can be resolved by adaptation in our framework.

5.4. Simulations

In this section, we report some simulation results on a Mandarin Chinese speech database. As addressed in Section 5.3, there are two component enabling technologies to support our bilateral user authentication framework for personalized mobile access. We report simulation results on two component technologies separately, and moreover, some speaker verification results enhanced by verbal information verification. Due to the limited space here, we report only the overall performance in all trials.

5.4.1 Database

For simulations, we use a Mandarin Chinese speech database of 50 people including 25 male and 25 female native speakers. The database consists of there sets for use in text-dependent and text-independent speaker verification as well as verbal information verification. In general, all the data in different sets are recorded in five sessions, labeled by $\mathbf{S}_1, \cdots, \mathbf{S}_5$, from one week to three weeks.

In the text-dependent data set, several fixed Chinese voice commands, e.g. ‘unlock’, ‘turn on’, and ‘hello’, are uttered three times in each session. For a voice command, thus, there are 15 fixed phrase utterances for each speaker in the set.

In the text-independent data set, we provide a set of conversation materials. In each session, a speaker in our database is asked to randomly select several sentences of over 30 seconds from the conversation set. As a result, there are the utterances of at least 30 seconds for each speaker in each session.

In the verbal information verification set, each speaker is asked to utter all the items in his/her private profiles registered in the system. Thus, there are five utterances for each item in the set. The use of five sessions is two-fold: enabling us to simulate multiple transactions and investigating the performance of our speaker-independent acoustic models, the kernel component of our verbal information verification system, in terms of voice aging. This data set will be used to test our Mandarin verbal information verification system.

As presented in Section 5.3.2, a set of speaker-independent HMMs are used for modeling Chinese acoustic units. For training the HMMs, we employ a benchmark Mandarin Chinese corpus, 863 corpus, in China. The population of this corpus is 200 people including 100 male and female speakers. For each speaker, there are a number of utterances ranging from 520 to 625 sentences elaborately selected from the database of the most famous Chinese newspaper – People Daily. Totally, all the utterances in this corpus correspond to up to 2185 sentences. Basically, almost all the phonetic information on Mandarin Chinese is covered by this corpus.

5.4.2 Speaker Verification

In this section, we report speaker verification results in terms of the text-dependent and text-independent data sets. For evaluating the performance of our methods, two different testing methods are used. one is to use *equal error* *rate* (EER), where the false rejection rate is equal to the false acceptance rate, without the need of a background model for decision-making. The other is to use a background model to yield real results, where we use *half total error rate* (HTER), defined as the average of the normalized false acceptance and false rejection rates, to evaluate the performance. For a specific speaker, speech data belonging to other people in the database are used as impostors' data during test.

Text-Dependent Experiments. For building an NFL speaker model, three utterances of a fixed phrase recorded in a specific session are used. Once one session is used for training, other four sessions are used for test. For reliability, we have performed five trials in the above way; five sessions are equally used as training and testing sets in five trials.

Now we present the acoustic analysis in text-dependent speaker verification. Before feature extraction, an utterance is pre-emphasized with the filter response $H(z) = 1 - 0.95z^{-1}$ and blocked into fixed-length frames. Each frame has 256 samples (2.56 ms) with 11.5 ms frame shift. The feature used is the statistical parameters of 19-order *Mel-scaled cepstrum coefficients* (MFCCs). The 19-order adaptive component weighted cepstrum coefficients [2] are superposed on MFCCs. Adaptive component weighted cepstrum is a robust feature to discriminate the speakers through emphasizing the formants of speakers.

Suppose that for an utterance belonging to speaker s, a set of N feature vectors, $X = \{\mathbf{x}_n^{(s)}\}_{n=1}^{N}$ where $\mathbf{x}_n = (x_{n,1}^{(s)}, \cdots, x_{n,19}^{(s)})^T$, are extracted by the above procedure. The mean and standard-deviation vectors,

$$\bar{\mathbf{x}}^{(s)} = (\bar{x}_1^{(s)}, \cdots, \bar{x}_{19}^{(s)})^T$$

and

$$\sigma_X^{(s)} = (\sigma_{X,1}^{(s)}, \cdots, \sigma_{X,19}^{(s)})^T,$$

are defined as

$$\bar{\mathbf{x}}^{(s)} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n^{(s)}$$

and

$$\sigma_{X,i}^{(s)} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(x_{n,i}^{(s)} - \bar{x}_i^{(s)}\right)^2}, \quad i = 1, \cdots, 19.$$

Thus, a new feature vector of this utterance, by integrating two statistics, is formed as

$$\hat{\mathbf{x}}^{(s)} = \left\{(\bar{\mathbf{x}}^{(s)}, \sigma_X^{(s)})\right\},$$

which is viewed to be a prototype of the NFL model corresponding to this speaker.

Table 5.1 The list of constitutions in five trails.

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Training Set	$\mathbf{S}_1$	$\mathbf{S}_2$	$\mathbf{S}_3$	$\mathbf{S}_4$	$\mathbf{S}_5$
Testing Set	$\mathbf{S}_2 \sim \mathbf{S}_5$	$\mathbf{S}_1$,$\mathbf{S}_3$,$\mathbf{S}_4$,$\mathbf{S}_5$	$\mathbf{S}_1$,$\mathbf{S}_2$,$\mathbf{S}_4$,$\mathbf{S}_5$	$\mathbf{S}_1$,$\mathbf{S}_2$,$\mathbf{S}_3$,$\mathbf{S}_5$	$\mathbf{S}_1 \sim \mathbf{S}_4$

On the basis of such a feature vector, each speaker's NFL model consists of three prototypes and, therefore, there are three feature lines resulting from three prototypes. In order to produce the pseudo-cohort models, each prototype is perturbed by two randomly produced vectors, $\delta\bar{\mathbf{x}}^{(s)}$ and $\delta\sigma_X^{(s)}$, respectively to form two pseudo-prototypes:

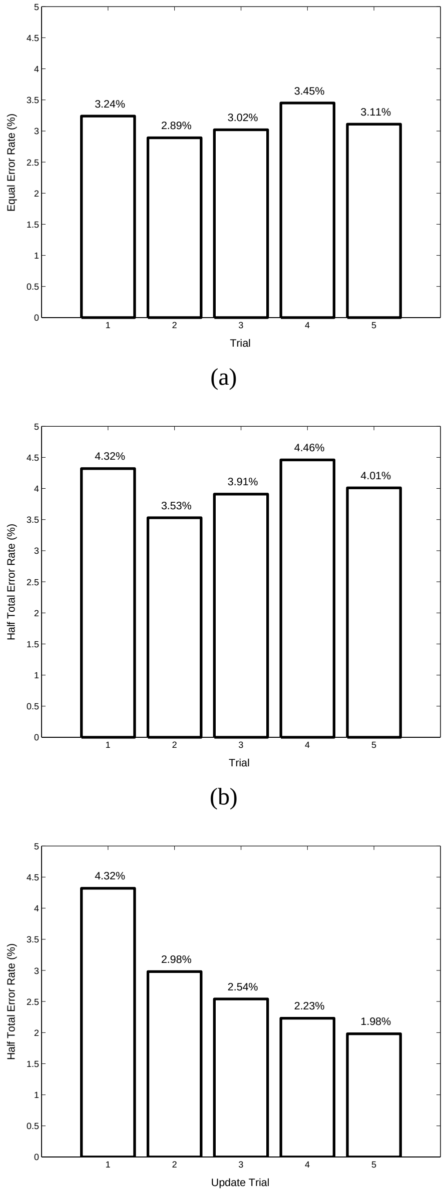
$$\hat{\mathbf{x}}_+^{(s)} = \left\{ (\bar{\mathbf{x}}^{(s)} + \delta\bar{\mathbf{x}}^{(s)}, \sigma_X^{(s)} + \delta\sigma_X^{(s)}) \right\}, \quad \hat{\mathbf{x}}_- = \left\{ (\bar{\mathbf{x}}^{(s)} - \delta\bar{\mathbf{x}}^{(s)}, \sigma_X^{(s)} - \delta\sigma_X)^{(s)} \right\}.$$

Accordingly, three pairs of pseudo-prototypes are formed to construct feature lines of pseudo cohort models corresponding to the speaker. It should be stated that the above perturbation is motivated by our previous studies on setting *a prior* threshold for speaker verification[7]. As a result, an acceptance/rejection decision is made through the competition between the speaker model and his/her pseudo-cohort models.

Figure 5.5 depicts simulation results in our experiments. Figure 5.5(a) shows the EERs in different trials, as listed in Table 5.1, by using only speaker models. By the pseudo-cohort models, we examine the performance of our system and show the HTERs in Figure 5.5(b). From Figures 5.5(a) and 5.5(b), the performance of our NFL-based system is reasonable for a fixed phrase of around 1.0 second. However, such error rates are still not acceptable for practical use. Therefore, we need to further improve the performance of our NFL system.

In fact, the error may result from miscellaneous mismatches, in particular, due to voice aging. Fortunately, new speech data should be always available as long as the handset is used. The availability of new data provides possibilities to update those prototypes in both the speaker and pseudo-cohort models. In order to alleviate the mismatch effects, we present a unsupervised on-line update method as follows. When the unknown utterance U, characterized by $\hat{\mathbf{x}}_U = \{(\bar{\mathbf{x}}_U, \sigma_U)\}$, corresponding to a fixed phrase is accepted by the speaker model no matter whether this decision is right or wrong, the system will update one of the previous prototypes in the speaker and the corresponding pseudo-cohort models. Assume that $\hat{\mathbf{x}}_{i^*}^{(s)}$ is the prototype satisfying the following condition:

$$i^* = \arg \min_{1 \leq i \leq 3} ||\hat{\mathbf{x}}_U - \hat{\mathbf{x}}_i^{(s)}||,$$



(a)

(b)

(c)

Figure 5.5. The performance of our NFL-based text-dependent speaker verification system installed in a handset. (a) Equal error rates in different trials. (b) The performance without update in different trials. (c) The performance with update in different trials.

Table 5.2 The list of constitutions in five update trails.

	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5
Training Set	$\mathbf{S}_1$	$\mathbf{S}_1, \mathbf{S}_2$	$\mathbf{S}_1 \sim \mathbf{S}_3$	$\mathbf{S}_1 \sim \mathbf{S}_4$	$\mathbf{S}_1 \sim \mathbf{S}_5$
Testing Set	$\mathbf{S}_2 \sim \mathbf{S}_5$	$\mathbf{S}_2 \sim \mathbf{S}_5$	$\mathbf{S}_3 \sim \mathbf{S}_5$	$\mathbf{S}_4, \mathbf{S}_5$	$\mathbf{S}_5$

where $||\cdot||$ is the Euclidean norm. Then, we replace this prototype with a new prototype, $\hat{\mathbf{x}}^{(s)}$, achieved by

$$\hat{\mathbf{x}}^{(s)} = \frac{\hat{\mathbf{x}}_U + \hat{\mathbf{x}}_{i^*}^{(s)}}{2}.$$

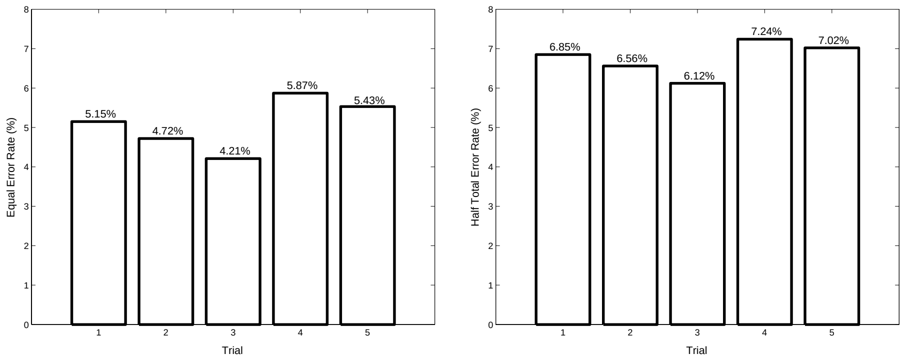
Accordingly, the corresponding prototypes, $\hat{\mathbf{x}}_{i^*,+}^{(s)}$ and $\hat{\mathbf{x}}_{i^*,-}^{(s)}$, in the pseudo-cohort model are updated based on the new prototype $\hat{\mathbf{x}}^{(s)}$.

To evaluate the performance of our system with the above on-line update mechanism, we conduct some experiments as listed in Table 5.2. As a result, Figure 5.5(c) illustrates the performance of our system with update. From Figure 5.5(c), it is observed that the performance in trail 1 is the same as the previous one shown in Figure 5.5(b) since there is no additional information available in trail 1. In trials 2-5, new data recorded in different sessions are available and, to some extent, the prototype update compensates for the mismatch due to voice aging. As illustrated in Figure 5.5(c), the error rate is lowered as more and more speech data recorded in different sessions are used for update. Here we emphasize that our update procedure performs autonomously and provides an alternative perspective towards the reduction of error rate in an adaptive way.

Text-Independent Experiments. For building a GMM-based speaker model, all the utterances recorded in a session are used. For reliability, we have performed five trials in the above way; five sessions are equally used as training and testing sets in five trials. In other words, one session is used for training and other four sessions are used for test in this trial.

Prior to training of a GMM speaker model, the acoustic analysis is performed as follows: a) pre-emphasizing with filter response $H(z) = 1 - 0.95z^{-1}$, b) 32ms Hamming windowing without overlapping, c) removing the silence and unvoiced part of speech in terms of short-term average energy, and d) extracting weighted 19-order Mel-scaled cepstral feature vector from each short-term frame.

In our simulations, the GMM of 32 Gaussian components is employed to characterize each speaker and a GMM of 512 Gaussian component is used to build a world background model. The GMM models are trained by the EM algorithm. All the data in the text-independent set are used to train the



(a) (b)

Figure 5.6. The performance of our GMM-based text-independent speaker verification system corresponding to speech segments of eight seconds. (a) Equal error rates in different trials. (b) Half total error rates in terms of the world background model.

world background model as done in the work [33]. As a result, all scores are normalized based on the world model prior to decision-making.

For decision-making, we adopt *a priori* speaker-dependent threshold setting method developed by ourselves [7]. In our method, we attempt to make a proper use of all the reliable statistics available. We believe that more reliable statistics provide more useful information, which might lead to a better threshold for decision-making. As a consequence, a speaker-dependent threshold is estimated by a linear combination of all the reliable statistics mentioned above:

$$T_S = b(\bar{\mu} + a\bar{\sigma}) + (1 - b)\mu, \tag{5.15}$$

where a and b are two speaker-independent parameters and optimized on the population of speakers used for building the world model. μ, $\bar{\mu}$, and $\bar{\sigma}$ are the statistics of normalized scores belonging to a speaker and the ensemble of impostors. Thus, Eg. (5.15) encodes the useful information conveyed by the reliable statistics, and the decision threshold becomes a monotonically increasing function of μ, $\bar{\mu}$, and $\bar{\sigma}$.

For test, we use the segment-based method presented in Section 5.3.1 to evaluate the performance of our system. Due to the limited space, we report only the results tested by speech segments of eight seconds. Figure 5.6 shows the performance of our GMM-based speaker verification system. Figure 5.6(a) depicts the EERs of our system in different trials, the same constitutions as used in the text-dependent experiments (c.f. Table 5.1), as by the use of only speaker models. In addition, the performance of our system by the use of the world background model is shown in Figure 5.6(b). From Figures 5.6(a) and 5.6(b), the performance of the GMM-based system is logic, which is consistent with other tests of a GMM-based text-dependent speaker verification system [28, 8].

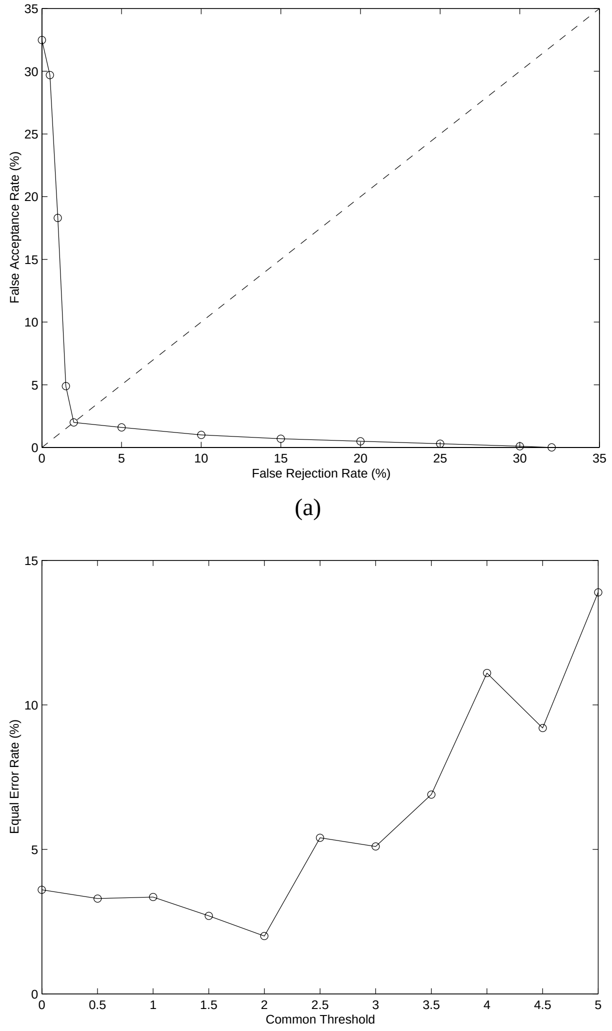
As illustrated in Figure 5.6, simulation results by only text-independent speaker verification system may lead to unacceptable error rates in practice. As pointed out in this paper, such a speaker verification system needs enhancing by incorporating other technologies, e.g. verbal information verification, to reduce error rates.

5.4.3 Mandarin Verbal Information Verification

In this section, we present simulation results of verbal information verification in terms of Mandarin Chinese. Moreover, we demonstrate how the verbal information verification system can enhance text-independent speaker verification.

Verbal Information Verification Experiments. To establish a Mandarin verbal information verification system, we use HMMs to model Mandarin Chinese acoustic units, INTIALs and FINALs. By a training and pruning procedure on the 863 corpus, totally, 467 tied models and 893 tied states form our speaker-independent acoustic modeling system. As a result, the performance of our acoustic modeling system reaches the accuracy rate of 71.8% in single syllable recognition. For test, a speaker is viewed as a true speaker only if the speaker's utterances are verified against his/her provide profile. On the other hand, this speaker will be considered as an impostor when the utterances are verified against other speakers' private profiles. For each true speaker, therefore, there are K utterances and $49K$ utterances from other 49 speakers as impostors (see Section 5.4.1), where K is the number of questions that a speaker is asked to answer. In the following experiments, the sequential utterance verification method presented in Section 5.3.2 is evaluated. Since there are five sessions, the overall performance in five trials are reported here. In our experiments, different thresholds, speaker-independent and context-dependent thresholds, are used to test our system.

Figure 5.7 illustrates the performance of our verbal information verification system by a single fixed threshold as three questions are asked. Figure 5.7(a) shows a receiver operating curve where the error rates in false rejection and false acceptance are achieved by changing the threshold value. Note that the above threshold is used for the utterance-level decision-making. In this experiment, we fix the subword threshold, $\theta = 2.0$, as defined in Eq. (5.14). As a result, our system reaches an EER of 2.0% in the experiment. Indeed, the performance of our system is also dependent upon the subword threshold, θ. For evaluating the performance of our system thoroughly, we also do an experiment by varying the subword threshold value in Eq. (5.14). As a consequence, the EERs of our system are depicted in Figure 5.7(b). It is observed from Figure 5.7(b)



(a)

(b)

Figure 5.7. The performance of our Mandarin verbal information verification system on a population of 50 people ($K = 3$, i.e., three questions are asked for each speaker). (a) Receiver operating curve at the subword threshold $\theta = 2.0$. (b) Equal error rates as the subword threshold θ varies from 0.0 to 5.0.

that the subword threshold θ results in the different performance though the utterance-level threshold is fixed.

Although the speaker-independent threshold used in our verbal information verification leads to the satisfactory performance in contrast to speaker verification, lower error rates are demanded for enhancing speaker verification. As a

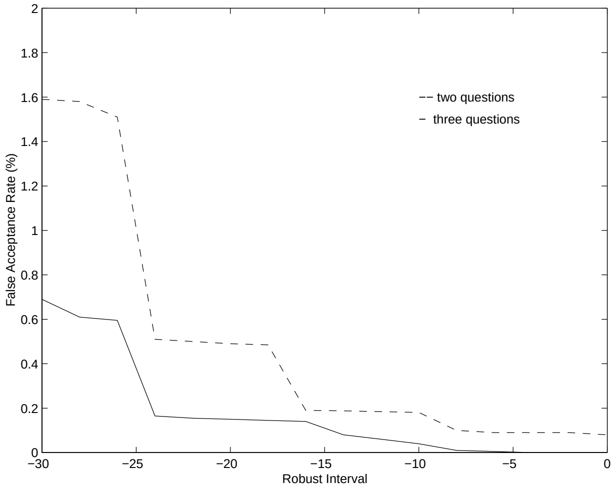


Figure 5.8. The performance of our Mandarin verbal information verification system on a population of 50 people when a speaker-dependent threshold is used.

result, we conduct another experiment by using a context-dependent threshold in our verbal information verification system. That is, different utterances are verified by different thresholds. Thus, the decision rule becomes

$$\mathrm{Acceptance}: U(i) \geq T(i), \;\; \mathrm{Rejection}: U(i) < T(i) \quad 1 \leq i \leq K.$$

Here $U(i)$ is the normalized confidence measure for each utterance and $T(i)$ is an original context-dependent utterance threshold. Since the variations in speech and environments lead to different testing scores for different speakers even for utterances of the same text. In order to characterize the variation and the system robustness, each utterance threshold is allowed to have a robust interval, τ. As a result, the threshold is adjusted as

$$T(i) = T(i) - \tau, \quad 0 \leq \tau \leq T(i). \tag{5.16}$$

In this circumstance, there are K thresholds associated with K questions. In our experiment, the thresholds are determined by first setting $T(i)$ such that the false rejection rate of our system is 0.0%. Then the thresholds are shifted to evaluate the false acceptance rate on different robust intervals τ as defined in Eq. (5.16). Figure 5.8 shows the relation between robust interval and false acceptance rates when two and three questions ($K = 2, 3$) are asked. As the robust interval varies, the evolutionary process of false acceptance rates are clearly shown in Figure 5.8. We can see that using two questions the system cannot reach an EER of 0.0% though some EERs are quite close to 0.0% . With three questions, our verbal information verification system yields an EER of 0.0% with 8.2%

robust interval. This three-question based result indicates that even though a true speaker's utterance scores are 8.2% lower than previous due to mismatch, the speaker still can be accepted while all the impostors in the database can be rejected correctly. Such a robust interval provides compensation for mismatch to ensure robust performance of our system, which makes the verbal information verification system qualify as a supervisor to enhance speaker verification.

Speaker Verification Enhanced by Verbal Information Verification. In our bilateral user authentication framework, the kernel technologies include text-independent speaker verification and verbal information verification. As presented in Sections 5.4.2 and 5.4.3, the text-independent speaker verification system does not yield satisfactory results for practical use, while the verbal information verification system reaches an EER of 0.0%. On the other hand, as pointed out in Section 5.1, text-independent speaker verification can work in a transparent way, while verbal information verification has to perform by an explicit query-based way. Therefore, a synergistic integration is to enhance text-independent speaker verification by verbal information verification as presented in Section 5.3.3.

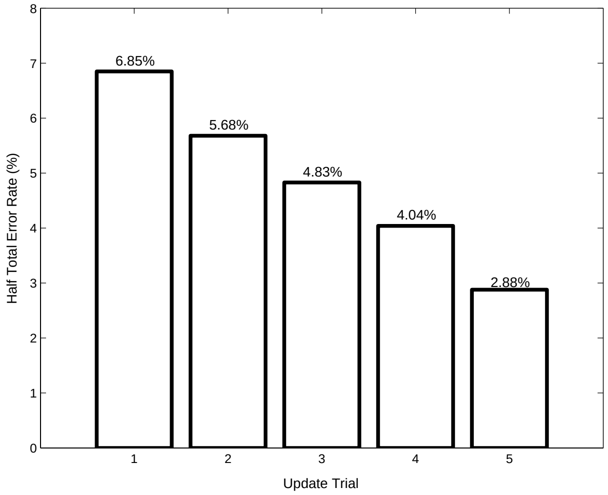


Figure 5.9. The performance of a text-independent speaker verification system enhanced by verbal information verification in terms of testing speech segments of eight seconds.

To evaluate the performance of the combination scheme presented in Section 5.3.3, we conduct an experiment to observe how the enhanced text-independent speaker verification system performs. Given a GMM-based speaker verification system, an unknown voice token is tested by this system. If the voice token is rejected, a verbal information verification procedure is started. If the user can pass the test of verbal information verification, it implies that the previous rejection caused by text-independent speaker verification is incorrect; that is, it is a false rejection. Thus, the user's utterances in this conversation is used to update the corresponding GMM-based speaker model. Similar to the work [33], we adopt an on-line EM algorithm to retrain a GMM for fast update. In our simulation, we use the whole session to update the current GMM-based speaker model once a speech segment in this session is incorrectly rejected. In our simulation, we use the GMMs trained on session 1 as a baseline system and the three-question based verbal information verification system with context-dependent thresholds for enhancement. As a result, we show the performance of the enhanced text-independent speaker verification system in Figure 5.9. For comparison, we depict the original performance of the baseline text-independent speaker verification system in trial 1. Trials 2-5 indicate that speech data in sessions 2-5 are sequentially used for test and update. From Figure 5.9, it is evident that the error rates are dramatically reduced during such a sequential update. Here, we emphasize that the above update is error-free since our verbal verification system reaches an EER of 0.0%, which provides an effective way towards reduction of error rates in speaker verification.

5.5. Conclusion

In this chapter, we have presented a bilateral user authentication to personalize mobile access in wireless environments where speaker verification, a biometric technology, and verbal information verification, a non-biometric technology, are integrated seamless to drive maximum synergy for user authentication. Enabling component speaker authentication technologies are described in terms of our own work. Simulations have been done separately for different enabling technologies and experimental results demonstrate that these enabling component technologies are ready for real application by building a complete bilateral user authentication system under our framework.

In our framework, we attempt to derive maximum synergy from the complementary capabilities of different speaker authentication technologies for security, easy-to-access, and friendly user interface in wireless environments. A salient feature is that such a framework can enhance the security and personalize mobile access in an adaptive way. Due to miscellaneous mismatches in voice characteristics, channels, and environments, our bilateral update strategies work efficiently towards continuous reduction of error rates in speaker authentication; the text-dependent speaker verification system in the client site is updated in an autonomous way, while the text-independent speaker verification system in the server site is updated in a supervised way by means of error-free verbal information verification. Indeed, several implementation issues for a complete bilateral user authentication system in real applications are not addressed in this chapter. Like component technologies presented in this chapter, these issues are not trivial at all and will be studied in the future development.

Acknowledgments

The partial work described here was done while the author worked at Peking University. The author is grateful to X.L. Li and T.Y. Wu for discussions and their help in simulations. The work described here was supported in part by an NSFC grant (60075017) and an MSR research grant on non-verbal speech analysis.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.

[2] K. T. Assaleh and R. J. Mammone. New LP-derived features for speaker identification. *IEEE Transactions on Speech and Audio Processing*, 2: 630–638, 1994.

[3] B. Atal. Automatic speaker recognition based on pitch contours. *Journal of Acoustics Society of America*, 52: 1687–1697, 1972.

[4] S. Barua. Authentication of cellular users through voice verification. In *Proceedings of IEEE International Conference on System, Man, and Cybernetics*, pages 420–425, 2000.

[5] F. Bimbot, M. Blomberg, L. Boves, D. Genoud, H.P. Hutter, C. Jaboulet, J. Koolwaaij, J. Lindberg, and J.B. Pierrot. An overview of the CAVE project research activities in speaker verification. *Speech Communication*, 31: 1437–1462, 2000.

[6] M. J. Carey and R. Auchenthaler. User validation for mobile telephones. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1093–1096, 2000.

[7] K. Chen. Towards better making a decision in speaker verification. *Pattern Recognition*, 35: (in press), 2002.

[8] K. Chen, L. Wang, and H. S. Chi. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification. *International Journal of Pattern Recognition and Artificial Intelligence*, 11: 417–445, 1997.

[9] K. Chen, T. Y. Wu, and H. J. Zhang. On the use of nearest feature line for speaker recognition. *Pattern Recognition Letters*, 23: (in press), 2002.

[10] K. Chen, D. H. Xie, and H. S. Chi. A modified HME for text-dependent speaker identification. *IEEE Transactions on Neural Networks*, 7: 1309–1313, 1996.

[11] R. V. Cox, C. A. Kamm, L. R. Rabiner, J. Schroeter, and J. G. Wilpon. Speech and language processing for next-millennium communication services. *Proceedings of The IEEE*, 88: 1314–1337, 2000.

[12] M. H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

[13] G. Doddington. Speaker recognition – Identifying people by their voice. *Proceedings of The IEEE*, 73: 1651–1664, 1985.

[14] G. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds. The NIST speaker recognition evaluation – Overview, methodology. *Speech Communication*, 31: 225–254, 2000.

[15] S. Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, 18: 859–872, 1997.

[16] A. K. Jain, R. Bolle, and S. Pankanti. *BIOMETRICS: Personal Identification in Network Society*. Kluwer Academic Publishers, 1999.

[17] B. H. Juang and S. Furui. Automatic recognition and understanding of spoken language – A first step toward natural human-machine communication. *Proceedings of The IEEE*, 88: 1142–1165, 2000.

[18] B. H. Juang and L. R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38: 1639–1641, 1990.

[19] L. S. Lee. Voice dictation of Mandarin Chinese. *IEEE Signal Processing Magazine*, 14: 63–101, 1997.

[20] L. S. Lee. Structural features of Chinese language – Why Chinese language processing is special and where we are. In *the International Symposium on Chinese Spoken Language Processing*, Singapore, 1998.

[21] Q. Li, B. H. Juang, C. H. Lee, Q. R. Zhu, and F. K. Soong. Recent advancements in automatic speaker authentication. *IEEE Robotics & Automation Magazine*, 6: 24–34, 1999.

[22] Q. Li, B. H. Juang, Q. Zhou, and C. H. Lee. Verbal information verification. In *Proceedings of EUROSPEECH*, pages 839–842, 1997.

[23] Q. Li, B. H. Juang, Q. Zhou, and C. H. Lee. Automatic verbal information verification for user authentication. *IEEE Transactions on Speech and Audio Processing*, 8: 585–596, 2000.

[24] X. L. Li and K. Chen. Mandarin Verbal Information Verification. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlando, 2002.

[25] J. M. Naik. Speaker recognition – A tutorial. *IEEE Communication Magazine*, 28: 42–48, 1990.

[26] J. Neyman and E. S. Pearson. On the use and interpretation of certain test criteria for purpose of statistical inference. *Biometrika*, 20A: 175–240, 1928.

[27] W. Reichl and W. Chou. Decision tree state tying based on segmental clustering for acoustic modeling. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 801–804, 1998.

[28] D. A. Reynolds. *A Gaussian mixture modeling approach to text-independent speaker identification*. Ph.D. Dissertation, Department of Electrical Engineering, Georgia Institute of Technology, 1992.

[29] A. E. Rosenberg. Automatic speaker verification: Review. *Proceedings of The IEEE*, 64: 475–487, 1976.

[30] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27: 43–49, 1978.

[31] O. Siohan, C. H. Lee, A. C. Surendran, and Q. Li. Background model design for flexible and portable speaker verification systems. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 825–828, 1999.

[32] D. Zhang. *Automated Biometrics: Technologies and Systems*. Kluwer Academic Publishers, 2000.

[33] Y. Y. Zhang, D. Zhang and X. Y. Zhu. A novel text-independent speaker verification method based on the global speaker model. *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 30: 598–602, 2000.