3) The registration function which assimilates the existing prototype model and the most recently acquired range image.
4) The NBV system which presents three separate methods for determining the NBV position from the current state of the model.
5) The graphical user interface with which the user can call the NBV system, view the images acquired at each iteration, review statistics pertaining to reconstruction, or examine the ideal or reconstructed model.
6) The application which reads in a reconstructed model file and outputs an IRIS Inventor format voxel rendering.
7) The IRIS Explorer module map which reads in a reconstructed model file and outputs an IRIS Inventor format surface rendering.

REFERENCES

[1] C. I. Connolly, "The determination of next best views," in *Proc. IEEE Int. Conf. Robotics and Automation*, 1985, pp. 432–435.
[2] J. Maver and R. Bajcsy, "Occlusions as a guide for planning the next view," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 417–433, May 1993.
[3] J. Maver, A. Leonardis, and F. Solina, "Planning the next view using the max-min principle," in *Computer Analysis of Images and Patterns*. New York: Springer-Verlag, Sept. 1993, pp. 543–547.
[4] R. Pito and R. Bajcsy, "A solution to the next best view problem for automated cad model acquisition of free-form objects using range cameras," in *Proc. SPIE*, vol. 2596, Oct. 1995, pp. 418–429.
[5] R. A. Jarvis, "A perspective for range finding techniques for computer vision," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 5, pp. 122–139, Apr. 1993.
[6] Y. Chen and G. Medioni, "Surface description of complex objects from multiple range images," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, 1994, pp. 153–158.
[7] C. H. Chien, Y. B. Sim, and J. K. Aggarwal, "Generation of volume/surface octree from range data," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, 1988, pp. 254–260.
[8] O. D. Faugeras and M. Hebert, "The representation, recognition, and positioning of 3-D shapes from range data," *Techn. 3-D Machine Percept.*, pp. 13–51, 1986.
[9] E. Bittar, S. Lavallé, and R. Szeliski, "A method for registering overlapping range images of arbitrarily shaped surfaces for 3-d object reconstruction," in *Proc. SPIE*, vol. 2059, 1993, pp. 384–395.
[10] H. Gagnon, M. Soucy, R. Bergevin, and D. Laurendeau, "Registration of multiple range views for automatic 3-d model building," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, June 1994, pp. 581–586.
[11] K. Higuchi, M. Herbert, and K. Ikeuchi, "Building 3-d models from unregistered range images," in *Proc. IEEE Int. Conf. Robotics and Automation*, May 1994, pp. 2248–2253.
[12] M. Rutishauser, M. Stricker, and M. Trobina, "Merging range images of arbitrarily shaped objects," *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, pp. 573–580, June 1994.
[13] M. Soucy and D. Laurendeau, "A general surface approach to the integration of a set of range views," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 344–358, Apr. 1995.
[14] B. A. Abidi, "Automatic sensor placement," in *Proc. SPIE*, vol. 2588, Oct. 1995, pp. 387–398.
[15] K. A. Taranabis, P. K. Allen, and R. Y. Tsai, "A survey of sensor planning in computer vision," *IEEE Trans. Robot. Automat.*, vol. 11, pp. 86–104, Feb. 1995.
[16] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*. Reading, MA: Addision-Wesley, 1992.

A Novel Text-Independent Speaker Verification Method Based on the Global Speaker Model

Yiying Zhang, David Zhang, and Xiaoyan Zhu

***Abstract*—This correspondence introduces a new text-independent speaker verification method, which is derived from the basic idea of pattern recognition that the discriminating ability of a classifier can be improved by removing the common information between classes. In looking for the common speech characteristics between a group of speakers, a global speaker model can be established. By subtracting the score acquired from this model, the conventional likelihood score is normalized with the consequence of more compact score distribution and lower equal error rates. Several experiments are carried out to demonstrate the effectiveness of the proposed method.**

***Index Terms*—Biometrics, equal error rate, global speaker model, speaker verification.**

I. INTRODUCTION

Speaker verification is the analysis of an utterance from an unknown speaker and its comparison with the model of the speaker whose identity is claimed with the verification result accepting or rejecting the claimed identity [1]. Generally speaking, speaker verification is a classifying problem of pattern recognition. It is expected that the separability between speakers will be more obvious if the common information between speakers is removed out by normalization.

Based on the above consideration, GSMSV (Speaker Verification based on the Global Speaker Model) method is proposed in this correspondence. In GSMSV, the global speaker model represents all of the common information contained in the speech of multiple speakers, and is utilized to normalize the likelihood score so that the difference between reference speakers and impostors is accentuated.

Research on speaker verification has been focused on speaker models [2], feature selection [3], and robust methods [4]. Higgins used a discriminate counting to verify the speaker [5], as well as likelihood score normalization methods [6]–[8], which are two likelihood score normalization methods by using impostor models. Since the method in [7] improves the speaker verification rate over the method in [6], it is used in this correspondence as a comparison method.

GSMSV is different from both the speaker verification methods with the conventional likelihood score (noted as CSV method in the following) [3] and the normalization method proposed in [6]. As we know, CSV method has some limitations, i.e., the loose distribution of likelihood scores leading to the vague boundaries between speakers and the burden to set a proper threshold, as well as the low system adaptability to protean input utterances. GSMSV method perfectly solves these limitations by employing the global speaker model to normalize the likelihood score.

The verification method proposed in [6] (called ASMSV in this correspondence) employs anti-speaker models to normalize the likelihood score, while GSMSV method establishes a global speaker model from all reference speakers to represent the common speech factors of different

Manuscript received January 22, 1999; revised June 20, 2000. The work was supported in part by the UGC (CRC) fund from the Hong Kong Government and the central fund from the Hong Kong Polytechnic University. This paper was recommended by Associate Editor M. S. Obaidat.

Y. Zhang is with the State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science, Tsinghua University, Beijing, China.

D. Zhang and X. Zhu are with the Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong.

Publisher Item Identifier S 1083-4427(00)07991-1.

speakers and to normalize the likelihood score. ASMSV method is faced with the conflict between the scale of the anti-speaker model (i.e., L, the number of speaker models included in an anti-speaker model) and verification speed. Furthermore, ASMSV method can not well distinguish the outside impostors, and the establishment of anti-speaker models is also difficult and time-consuming. While in GSMSV method, the global speaker model is easy to obtain and the verification speed is very fast.

II. GSMSV Method

A. Definitions and Notations

Likelihood score in speaker recognition can be generally defined as the matching score of a test utterance to a specific speaker model. Given an input utterance Y, and a speaker model λ, the likelihood score is the probability of λ produces Y, i.e., $P(Y|\lambda)$.

Given N reference speakers, whose models are $\lambda_1, \cdots, \lambda_i, \cdots, \lambda_N$, in which λ_i is obtained by maximizing the likelihood score $P(Y_i|\lambda_i)$, and Y_i is the training data of reference speaker i, the global speaker model, $\lambda_{\rm GSM}$, is established in GSMSV method besides the N reference speaker models, by maximizing $\prod_{i=1}^{N} P(Y_i|\lambda_{\rm GSM})$.

$\lambda_{\rm GSM}$ contains not only the common speech characteristics of multiple speakers, but also the environmental features related to the speaking background. According to the principal idea of pattern recognition that removing the common information is helpful to improve a classifier's discriminating ability, it is anticipated that the differences between speakers will be emphasized if the common speech characteristics are obliterated from speech. Based on this consideration, GSMSV method is designed as follows:

Let $S^{(i)}_{\rm GSM}(Y)$ be the normalized likelihood score for an input utterance, Y, claimed to be uttered by the ith reference speaker, we can get $S^{(i)}_{\rm GSM}(Y) = P(Y|\lambda_i) - P(Y|\lambda_{\rm GSM})$. By subtracting the score obtained from $\lambda_{\rm GSM}$, the common information of both pronunciation characteristics and environmental features is obliterated. As a result, the interference of unimportant factors is avoided, and the differences between speakers are brought into prominence. Therefore, the decision rule can be defined as

$$S^{(i)}_{\rm GSM}(Y) \begin{cases} > \eta, & \text{accept the claim to reference speaker } i \\ \le \eta, & \text{reject the claim to reference speaker } i \end{cases} \tag{1}$$

where η is a threshold. To avoid overflow in computation, logarithm likelihood score is utilized. Let $LS^{(i)}_{\rm GSM}(Y)$ be the normalized logarithm likelihood score for an input utterance, Y, claimed to be uttered by the ith reference speaker, we can get $LS^{(i)}_{\rm GSM}(Y) = \log P(Y|\lambda_i) - \log P(Y|\lambda_{\rm GSM})$. Thus the decision rule can be represented as

$$LS^{(i)}_{\rm GSM}(Y) \begin{cases} > \eta', & \text{accept the claim to reference speaker } i \\ \le \eta', & \text{reject the claim to reference speaker } i \end{cases} \tag{2}$$

where η' is a threshold.

To further improve the system adaptability and alleviate the influence of speaking rate, the logarithm likelihood score is normalized again by duration as in the following:

$$\frac{LS^{(i)}_{\rm GSM}(Y)}{T_Y} \begin{cases} > \eta'', & \text{accept the claim to reference speaker } i \\ \le \eta'', & \text{reject the claim to reference speaker } i \end{cases} \tag{3}$$

where T_Y is the number of input speech frames, and η'' is a threshold.

B. General Estimation of the Global Speaker Model

In this correspondence, the speaker model is Gaussian mixture model (GMM). Its distribution of the training speech data in the acoustic space for each reference speaker is represented by mixture Gaussian probability density functions, which is similar to "semi-continuous" probability distribution or "tied mixture" technique for representing speech segments in hidden Markov based speech recognition [8]. The parameters of a GMM can be represented as: $\lambda = ((c_1, \mu_1, \Sigma_1), \cdots, (c_k, \mu_k, \Sigma_k), \cdots, (c_M, \mu_M, \Sigma_M))$, in which μ_k and Σ_k are mean vector and covariance matrix for the kth Gaussian density function, respectively; c_k is the corresponding kth weight; and M is the number of mixture components. Let $Y = \{y_1, \cdots, y_k, \cdots, y_T\}$ be the sequence of feature vectors for an input utterance, thus its likelihood score produced by λ is obtained as

$$P(y_k|\lambda) = \sum_{m=1}^{M} c_m \cdot \frac{1}{(\sqrt{2\pi})^{D/2} \cdot (|\Sigma_m|)^{1/2}} \cdot \exp\left(-\frac{1}{2}(y_k - \mu_m)^T \Sigma_m^{-1} (y_k - \mu_m)\right) \tag{4}$$

and $P(Y|\lambda) = \prod_{k=1}^{T} P(y_k|\lambda)$, where D is the dimensionality of feature vectors.

Assume the current verification system has N users, whose training data is represented as $Y_i = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_k^{(i)}, \cdots, y_{T(i)}^{(i)}\}$ $(i = 1, 2, \cdots, N)$, after being transformed to feature vectors, in which i denotes the ith speaker and $T(i)$ denotes the total number of feature vectors for the ith speaker. The training data for a new system user, the $(N+1)$th reference speaker, is noted as

$$Y_{N+1} = \{y_1^{(N+1)}, y_2^{(N+1)}, \cdots, y_k^{(N+1)}, \cdots, y_{T(N+1)}^{(N+1)}\}.$$

Let the parameters of $\lambda_{\rm GSM}$ be

$$\lambda_{\rm GSM} = ((c_1^{\rm GSM} \mu_1^{\rm GSM} \Sigma_1^{\rm GSM}) \cdots, (c_k^{\rm GSM} \mu_k^{\rm GSM} \Sigma_k^{\rm GSM}) \cdots, (c_M^{\rm GSM} \mu_M^{\rm GSM} \Sigma_M^{\rm GSM}))$$

in which $c_k^{\rm GSM}$ is the weight of the kth Gaussian density function; $\mu_k^{\rm GSM}$ and $\Sigma_k^{\rm GSM}$ are the corresponding mean vector and covariance matrix, respectively. In the general re-estimation method, the parameters of $\lambda_{\rm GSM}$ are obtained by maximizing $\prod_{i=1}^{N+1} P(Y_i|\lambda_{\rm GSM})$ with Maximum Likelihood criterion [9], which is an iterative procedure starting from the initial values set by Segmental K-means procedure [10]. Thus the re-estimation formulas for $\lambda_{\rm GSM}$ are as follows:

$$\hat{c}_j^{\rm GSM} = \frac{\sum_{n=1}^{N+1} \sum_{t=1}^{T(n)} \theta_j^{(n)}(t)}{\sum_{n=1}^{N+1} \sum_{t=1}^{T(n)} \alpha_t^{(n)} \cdot \beta_t^{(n)}} \qquad j = 1, 2, \cdots, M \tag{5}$$

$$\theta_j^{(n)}(t) = \begin{cases} c_j^{\rm GSM} p_j[y_1^{(n)}] \beta_1^{(n)} & t = 1 \\ c_j^{\rm GSM} p_j[y_t^{(n)}] \alpha_{t-1}^{(n)} \beta_t^{(n)} & t = 2, 3, \cdots, T(n) \end{cases} \tag{6}$$

$$\alpha_t^{(n)} = \begin{cases} p[y_t^{(n)}] \alpha_{t-1}^{(n)} & t = 2, 3, \cdots, T(n) \\ p[y_1^{(n)}] & t = 1 \end{cases} \tag{7}$$

$$\beta_t^{(n)} = \begin{cases} p[y_{t+1}^{(n)}] \beta_{t+1}^{(n)} & t = 1, 2, \cdots, (T(n) - 1) \\ 1 & t = T(n) \end{cases} \tag{8}$$

$$\hat{\mu}_j^{\rm GSM} = \frac{\sum_{n=1}^{N+1} \sum_{t=1}^{T(n)} \theta_j^{(n)}(t) y_t^{(n)}}{\sum_{n=1}^{N+1} \sum_{t=1}^{T(n)} \theta_j^{(n)}(t)} \qquad j = 1, 2, \cdots, M \tag{9}$$

$$\hat{\Sigma}_j^{\mathrm{GSM}} = \frac{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)\cdot(y_t^{(n)}-\hat{\mu}_j^{\mathrm{GSM}})(y_t^{(n)}-\hat{\mu}_j^{\mathrm{GSM}})^T}{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)} \quad j=1,2,\cdots,M \tag{10}$$

where $\hat{c}_j^{\mathrm{GSM}}$, $\hat{\mu}_j^{\mathrm{GSM}}$, and $\hat{\Sigma}_j^{\mathrm{GSM}}$ are the latest values; c_j^{GSM}, μ_j^{GSM}, and Σ_j^{GSM} are their corresponding values of the last iteration. In (5)–(10), $p[y_t^{(n)}]$ is same as $p[y_t^{(n)}|\lambda_{\mathrm{GSM}}]$, and

$$p_j[y_t^{(n)}] = \frac{1}{(\sqrt{2\pi})^{D/2}\cdot(|\Sigma_j^{\mathrm{GSM}}|)^{1/2}} \cdot\exp\left(-\frac{1}{2}(y_t^{(n)}-\mu_j^{\mathrm{GSM}})^T \cdot(\Sigma_j^{\mathrm{GSM}})^{-1}(y_t^{(n)}-\mu_j^{\mathrm{GSM}})\right).$$

III. Real-Time Application

In GSMSV method, the global speaker model is a critical factor directly influencing the system performance and practical application. Therefore, the method of establishing the global speaker model is an important issue worthy of discussing. In Section II-B, the general re-estimation method, which obtains the approximately optimal parameters of λ_{GSM}, has been introduced. This section is to emphasize on its limitation and to present an adaptive method that can quickly adapt the parameters of λ_{GSM} to a new user without decreasing the verification rate for the old ones.

Since λ_{GSM} is obtained by using all the training data of current users, the training procedure takes a long time, especially when the system has a large number of users. In our experiments, when the current system has 100 users, a new registration needs about 40 min. This is unacceptable and intolerable for real-time applications.

There are two main causes leading to the slow estimation of the parameters of λ_{GSM}. One is that the parameter estimation is an iterative procedure. The other is that the initial values of the iterative procedure are set by time-consuming K-means procedure. So the adaptive estimation focuses on these two factors, updating the parameters in a one-shot step. The initial values are set as those modified by the last registration. The adaptive re-estimation formulas are as follows: (see (11) and (12) at the bottom of the page)

$$\hat{\Sigma}_j^{\mathrm{GSM}} = \frac{(1-\rho)\cdot A+\rho\cdot B}{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)} \quad j=1,2,\cdots,M \tag{13}$$

$$A = \sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)\cdot(y_t^{(n)}-\hat{\mu}_j^{\mathrm{GSM}})\cdot(y_t^{(n)}-\hat{\mu}_j^{\mathrm{GSM}})^T \tag{14}$$

$$B = \sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)\cdot(y_t^{(N+1)}-\hat{\mu}_j^{\mathrm{GSM}}) \cdot(y_t^{(N+1)}-\hat{\mu}_j^{\mathrm{GSM}})^T. \tag{15}$$

In (11)–(15), $\theta_j^{(n)}(t)$, $\alpha_t^{(n)}$ and $\beta_t^{(n)}$ are computed as (8)–(10).

In adaptive re-estimation, a weighting coefficient ρ $(0<\rho<1)$ is introduced to measure the contribution of the new registration speech to updating the global speaker model. The greater the value of ρ, the more the contribution owing to the new training data. Since the adaptive re-estimation procedure starts from the last modified parameter values, the setting of ρ will determine the verification performance after the system scale is augmented. Without ρ or it is too small), the system would not adapt to the new user. If ρ is too large, the global speaker model is changed exceedingly to accommodate the new user, but it may not be applicable to the old ones.

ρ may be set as a specific value according to experimental experience or the change of the number of reference speakers. For example, ρ can be decided by $\rho=\epsilon\cdot R(N_{\mathrm{users}})$, in which ϵ is a coefficient showing the greatest portion of the new speaker's contribution to all the old speakers' contribution, and $R(N_{\mathrm{users}})$ is a function whose values increase with the number of valid users, N_{users}, and whose limitation value is 1, e.g., $R(N_{\mathrm{users}})$ can be a sigmoid function.

IV. Experiments

A. Database and Experimental Settings

Data used in the following experiments come from a standard Mandarin speech database *863Bag* provided by the *State Education Commission* of China. Speech data of 25 females and 25 males is used. Each person uttered 50 sentences, 15 of which are used as the training data, and the other 35 sentences are used as the test data. Each test is on one sentence. The average duration of training data for each speaker is about 60 s, and that of each test utterance is about 3.5 s.

Fifteen females and 15 males are regarded as reference speakers. Tests on their data consist of closed set test, in which the speech of one reference speaker makes up the disguised utterance to other reference speakers. Tests on data of 20 other speakers (ten females and ten males) who are regarded as outside impostors constitute open set test.

Equal error rate is used to measure the performance of different speaker verification methods. The *posterior* equal error rate is a convenient measure of the degree of separation between true and false speaker scores and, therefore, a useful predictor of speaker verification performance. In the following experiments, serials of values on

$$\hat{c}_j^{\mathrm{GSM}} = \frac{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)}{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\alpha_t^{(n)}\cdot\beta_t^{(n)}+\rho\cdot\sum_{t=1}^{T(N+1)}\alpha_t^{(N+1)}\cdot\beta_t^{(N+1)}} \quad j=1,2,\cdots,M \tag{11}$$

$$\hat{\mu}_j^{\mathrm{GSM}} = \frac{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)y_t^{(n)}+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)y_t^{(N+1)}}{(1-\rho)\cdot\sum_{n=1}^{N}\sum_{t=1}^{T(n)}\theta_j^{(n)}(t)+\rho\cdot\sum_{t=1}^{T(N+1)}\theta_j^{(N+1)}(t)} \quad j=1,2,\cdots,M \tag{12}$$

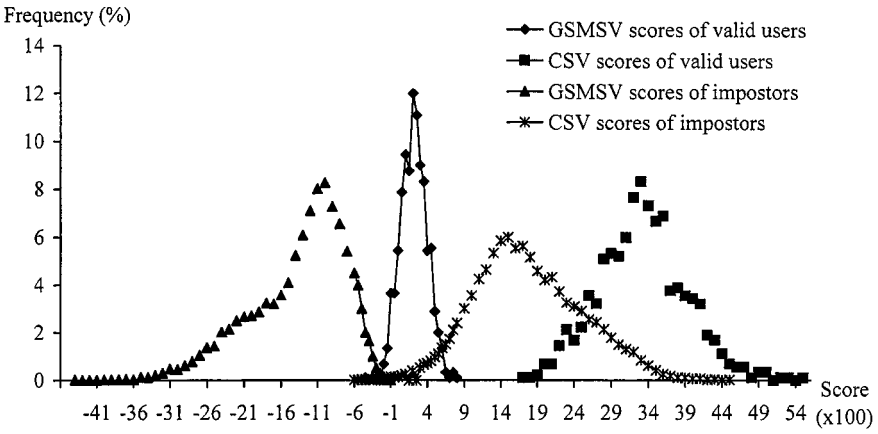

Fig. 1. Likelihood score histograms of the closed set tests for CSV and GSMSV methods.

TABLE I
STATISTICAL ANALYSIS RESULTS OF THE LIKELIHOOD SCORES FOR CSV METHOD AND GSMSV METHOD

	Valid speakers		Impostors		Likelihood score difference between valid users and impostors
Method	Mean	Variance	Mean	Variance	d
CSV	3354.79	584.36	1762.85	730.89	276.69
GSMSV	285.32	188.13	-1312.71	641.47	768.43

TABLE II
PERFORMANCE COMPARISON (EQUAL ERROR RATES AND VERIFICATION SPEED) OF DIFFERENT METHODS

Method	Closed set test (%)	Open set test (%)	Average verification speed (s)
CSV	6.19	1.69	0.57
ASMSV (L=29)	0.19	1.06	17.26
GSMSV	0.59	0.51	1.15
ASMSV (L=1)	7.80	2.16	1.15

decision boundary are tried to make the error rate of false rejection is equal to that of false acceptance, thus the equal error rate is found and used as a measure for comparison.

The speech signal was sampled at a rate of 8 kHz, segmented into 32 ms overlapping frames with a 16 ms shift, and pre-emphasized. A feature vector consists of 16 cepstrum coefficients acquired from auto-relation analysis, 16 dynamic cepstrum coefficients and a dynamic energy [11]. GMM speaker model has 64 mixtures. ρ is set to be a specific value, 0.2. The computer used for experiments is P-II 233.

B. Statistical Analysis

In this experiment, the statistical likelihood scores of both CSV and GSMSV methods are analyzed and compared. The likelihood scores of closed set test are recorded and the corresponding histograms are shown in Fig. 1. The statistical results are also listed in Table I, in which d is the likelihood score difference between valid users and impostors.

The following interesting observations can be obtained:

1) For the speech of either valid users or impostors, the variance of GSMSV likelihood scores is much smaller than that of CSV method. This illustrates that GSMSV makes the distribution of the likelihood score more compact.
2) The difference between the likelihood scores of GSMSV valid users and impostors is greater than that of CSV method. It demonstrates that GSMSV enlarges the distance between valid users and impostors, so its distinguishing ability is more powerful.
3) The GSMSV likelihood score overlap between valid users and impostors is smaller than that of CSV method, therefore the boundary between valid users and impostors is more explicit and the threshold can be more conveniently set by GSMSV.

C. Comparison of Different Methods

ASMSV method has the lowest equal error rates when an anti-speaker model consists of all of other reference speakers [7], thus in this experiment the value of L is set to be 29. For the sake of comparison convenience, the results of the case $(L = 1)$ are also given. Table II lists the equal error rates of CSV, ASMSV and GSMSV methods.

In Table II, the equal error rates of both ASMSV $(L = 29)$ and GSMSV methods are all significantly lower than those of CSV method. This shows the necessity to normalize the likelihood score. For closed set test, the equal error rate of GSMSV is higher than that of ASMSV, but for open set test the equal error rate of GSMSV is much lower. Besides, it should be noted that the equal error rates under the case $(L = 29)$ are the best results that ASMSV can reach.

Table II also lists the average time needed by each method to verify an input utterance. It costs ASMSV over 17 s to verify an utterance, while GSMSV spends only about 1 s. If ASMSV spends 1 s to verify an utterance, its equal error rates are much higher than the corresponding values of GSMSV method.

The above experiments show that GSMSV method has the advantages on both lower equal error rates and faster verification procedure. Either CSV method or ASMSV method can not keep low equal error rates and fast verification speed at the same time.

D. Experimental Results of General and Adaptive Gsmsv

A serial of experiments is performed on different number of reference speakers. Experiments start from two reference speakers (one female, one male). And then in the following experiments, one female and one male are added each time. In each experiment, λ_{GSM} is updated two times, by firstly using the training data of the new female user, and then using that of the new male user. After these modifications, λ_{GSM} is used for verification tests. The test results are depicted in Figs. 2 and 3, respectively.

The following observations can be obtained. For both closed set test and open set test, GSMSV with either the general re-estimation or the adaptive re-estimation method has much lower equal error rates than CSV method. The equal error rates of the adaptive re-estimation approximate to those of the general re-estimation.

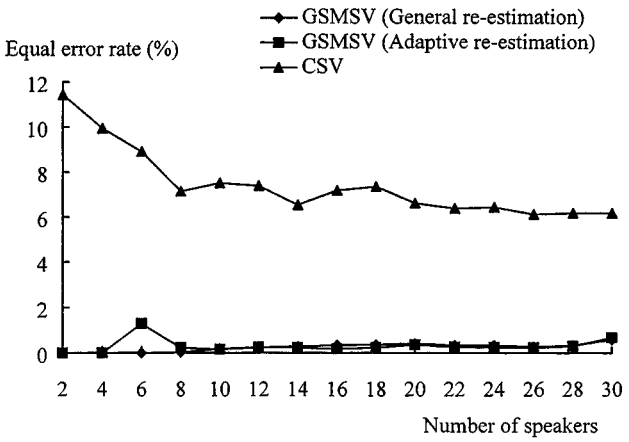

Fig. 2. GSMSV test results of the closed set test with different re-estimation methods for updating the global speaker model.

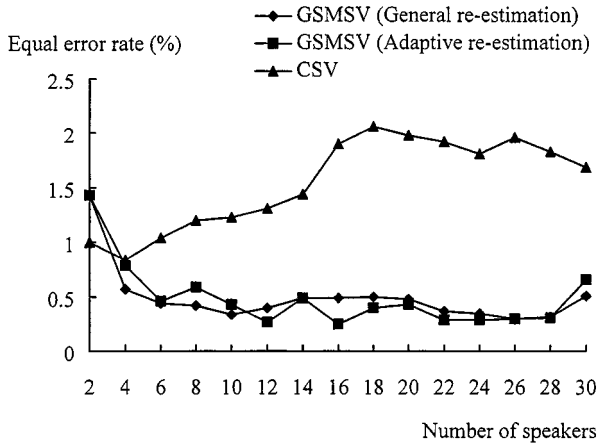

Fig. 3. GSMSV test results of the open set test with different re-estimation methods for updating the global speaker model.

When two new users (one female and one male) are added into the system, the training time for GSMSV is recorded and depicted in Fig. 4. The training time for the adaptive re-estimation increases a little with extension of the system, while that for the general re-estimation increases proportionally and significantly.

The effectiveness and practicability of the adaptive re-estimation method has been fully illustrated by these experiments. Compared to the general method, the adaptive one decreases the registration time significantly without increasing the equal error rates.

V. Conclusion

A novel speaker verification method, GSMSV, is proposed in this correspondence. GSMSV has the following characteristics:

2) The separation between speakers is large and explicit.
3) The distinguishing ability of the system is powerful.
4) Verification speed is fast.
5) It is adaptable to speaking speed.

As a result of the proposed likelihood score normalization employing the global speaker model, the equal error rates are decreased with a fast

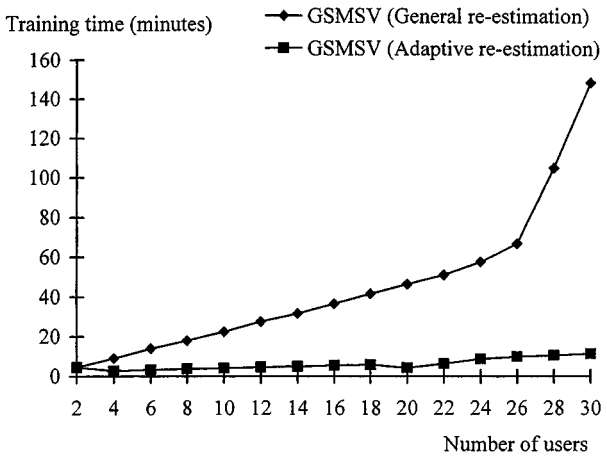

Fig. 4. Training time for GSMSV with different re-estimation methods for updating the global speaker model when two new users register into the verification system.

verification speed. Compared with ASMSV ($L = 29$), GSMSV speeds up the verification procedure with decrease on the verification time by 93%. By contrast to CSV method, GSMSV decreases the equal error rates by 90% and 70% for closed set test and open set test, respectively.

In order to apply GSMSV method to real-time systems, an adaptive re-estimation approach to updating the global speaker model is suggested to shorten the waiting time for a new user. When the system has 30 users, registration is accelerated 12 times.

References

[1] J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, pp. 1437–1462, Sept. 1997.
[2] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 72–83, Jan. 1995.
[3] A. Haydar, M. Demirekler, and M. K. Yurtseven, "Feature selection using genetic algorithm and its application to speaker verification," *Electron. Lett.*, vol. 34, no. 15, pp. 1457–1459, July 1998.
[4] J. M. Richard, X. Y. Zhang, and R. Ramachandran, "Robust speaker recognition—A feature-based approach," *IEEE Signal Processing Mag.*, pp. 58–71, Sept. 1996.
[5] Higgins and L. Bahler, "Text-independent speaker verification by discriminant counting," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, May 1991, pp. 405–408.
[6] E. Rosenberg, J. Delong, C. H. Lee, B. H. Juang, and F. K. Soong, "The use of cohort normalized scores for speaker recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 1, Oct. 1992, pp. 599–602.
[7] C. S. Liu, H. C. Wang, and C. H. Lee, "Speaker verification using normalized log-likelihood score," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 57–60, Jan. 1996.
[8] X. D. Huang and M. A. Jack, "Semi-continuous hidden Markov models for speech signals," *Computer Speech and Language*, vol. 3, pp. 239–251, 1989.
[9] L. A. Liporace, "Maximum likelihood estimation for multivariate observations of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 729–734, Sept. 1982.
[10] L. R. Rabiner, "Recognition of isolated digits using hidden Markov mjodels with continuous mixture densities," *AT&T Tech. J.*, vol. 64, no. 6, pp. 1211–1222, 1985.
[11] C. H. Lee, "On robust linear prediction of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, pp. 642–650, 1988.