A NOVEL TEXT-INDEPENDENT SPEAKER VERIFICATION METHOD USING THE GLOBAL SPEAKER MODEL

Yiying Zhang, Xiaoyan Zhu, IEEE Member

State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing 100084
Email: zxy-dcs@mail.tsinghua.edu.cn

ABSTRACT

In this paper a new text-independent speaker verification method is proposed based on likelihood score normalization and the global speaker model, which is established to represent the universal features of speech and environment, and to normalize the likelihood score. As a result the equal error rates are decreased significantly, verification procedure is accelerated and system adaptability is improved. Two possible ways of establishing the global speaker model, one of which can meet the real-time requirement, are also suggested and discussed. Experiments demonstrate the effectiveness of this novel verification method and its improvement over the conventional method and other normalization methods.

1. INTRODUCTION

In the conventional speaker verification method (noted as CSV in the following), the decision rule of acceptance or rejection is based on the score of a test utterance for the claimed speaker and a predefined threshold [1]. The CSV method has three limitations. The first is the loose distribution of likelihood score, which leads to vague boundaries between speakers. The second is the burden to set a proper threshold, which is a direct consequence of scattered likelihood score. The third is low system adaptability to protean input utterances with different duration and different content.

The proposed text-independent speaker verification method using the global speaker model (called as GSMSV method in this paper) is based on our previous work [2] and aims at the limitations of the conventional method. The global speaker model is established to represent all of the common information between speakers, such as the pronunciation characteristics, background noises, common features of texts with different contents. The likelihood score is normalized by the score produced by the global speaker model. As a result, the score distribution is much more concentrated.

The verification method proposed in [3] is also a method of likelihood score normalization, which is based on anti-speaker model (called ASMSV in this paper). Although ASMSV method can also significantly improve the performance of CSV method, it faces the conflict between the number of speaker models included in an anti-speaker model (represented as *L* in [3]) and verification speed. With increase in the value of *L*, the verification speed becomes slower and slower. On the other hand, if *L* is too small, the equal error rates are very high. In addition, establishing anti-speaker models is a time-consuming procedure. GSMSV method avoids the dilemma facing ASMSV method. The global speaker model is easy to obtain and the verification speed is very fast.

2. THE PROPOSED METHOD

2.1. Method Description

Given N reference speakers, whose models are $\lambda_1, \cdots, \lambda_i, \cdots, \lambda_N$ respectively, in which λ_i is obtained by maximizing likelihood score $P(Y_i \mid \lambda_i)$, and Y_i is the training data of reference speaker *i*. In GSMSV method, an extra model, global speaker model λ_{GSM} is added. λ_{GSM} is acquired by maximizing $\prod_{i=1}^{N} P(Y_i \mid \lambda_{GSM})$, i.e., the training data for λ_{GSM} includes the data of all reference speakers. Thus there are totally $N+1$ speaker models, in which λ_{GSM} represents the universal speech characteristics of multiple speakers.

Since λ_{GSM} is acquired from the training data of all reference speakers, it includes the common information related to pronunciation, speaking environment and text contents. If these information is removed out from speech, the differences between speakers will be emphasized. Thus GSMSV utilizes λ_{GSM} to normalize likelihood score and to exclude all common information contained in speech.

Let S_{GSM} be the normalized likelihood score of GSMSV method. To an input utterance *X*, it is computed as the following.

$$S_{GSM} = P(X|\lambda_i) - P(X|\lambda_{GSM})$$

By subtracting the score produced by λ_{GSM}, the common information is obliterated from speech. As a result, the interference of unimportant factors is avoided and the distinction between different speakers is clearer. Therefore the decision rule for GSMSV is:

$$P(X \mid \lambda_i) - P(X \mid \lambda_{GSM}) \begin{cases} > \eta, & \textit{Accept the claim} \\ \le \eta, & \textit{Reject the claim} \end{cases} \quad (1)$$

in which η is a threshold.

One point should be noted. Although λ_{GSM} is obtained from the training data of reference speakers, it is a universal speaker model, not only representing common characteristics of reference speakers, but also embodying the common information of outside impostors. Therefore theoretically GSMSV has a powerful ability to distinguish both reference speakers and outside impostors.

In order to avoid overflow in computation, log likelihood score is utilized and thus the decision rule becomes

$$logP(X \mid \lambda_i) - logP(X \mid \lambda_{GSM}) \begin{cases} > \eta', & Accept\ the\ claim \\ \le \eta', & Reject\ the\ claim \end{cases} \quad (2)$$

in which η' is a threshold.

To further improve the adaptability of text-independent verification system and alleviate the influence of utterance duration, likelihood score is further normalized by the duration of test utterance. Thus decision rule (2) is improved by formula (3), in which T_X is the number of frames and η'' is a threshold.

$$\frac{logP(X \mid \lambda_i) - logP(X \mid \lambda_{GSM})}{T_X} \begin{cases} > \eta'', & Accept\ the\ claim \\ \le \eta'', & Reject\ the\ claim \end{cases} \quad (3)$$

Compared with CSV, GSMSV improves the distribution of likelihood score, making it more centered and is more adaptive since it removes the influence of speaking speed and thus further improve system adaptability. In addition, setting threshold is much easier owing to more compact score distribution.

Compared with ASMSV method, establishing λ_{GSM} does not need the procedure to find the typical impostors for a reference speaker. It is obtained as to create the usual speaker model, and the difference only lies in the training data used. Furthermore two calculations of likelihood scores during verification phase are also very fast and can satisfy real-time verification requirements.

2.2. Comparison with Other Methods

One of the popular methods for normalizing likelihood score is to establish Anti-Speaker Model (noted as ASM) [3]. The ASM of a reference speaker is the subset of reference speaker models. ASM represents the speech characteristics of impostors. Thus speaker space is divided into two sub-spaces, claimed reference speaker and his impostors.

Let λ_i and λ_i' represent the speaker model of certain reference speaker *i* and his impostors respectively. Thus

$$\lambda_i' = \{\lambda_{s(1)}, \lambda_{s(2)}, \cdots, \lambda_{s(L)}\}$$
$$s(k) \in [1..N], s(k) \neq i, \quad k = 1, 2, \cdots, L$$

in which N is the number of reference speakers, L is the number of reference speakers in the ASM. λ_i and λ_i' are obtained by maximizing $P(Y \mid \lambda_i)$ and $P(Y' \mid \lambda_i')$ using Maximum Likelihood criterion, in which Y and Y' are training data of reference speaker and impostors. Given test data X, the decision rule for deciding whether X is uttered by the *i*-th reference speaker is:

$$\frac{P(X \mid \lambda_i)}{P(X \mid \lambda_i')} \begin{cases} > \eta, & Accept\ the\ claim \\ \le \eta, & Reject\ the\ claim \end{cases}$$

From the above description, we can see that there exists great differences between CSV, ASMSV and GSMSV methods. As a summary, their differences are listed in Table 1.

Method	CSV	ASMSV	GSMSV
Training Procedure	Train N models	Train N models and establish L ASM models	Train N models and 1 global speaker model
Whether to Normalize Score	No	Yes	Yes
Score Calculations for Verification	1 time	L+1 times	2 times

Table 1: Differences between three methods.

3. ESTABLISHMENT OF THE GLOBAL SPEAKER MODEL

Since the global speaker model is critical to the performance of verification system, the parameter estimation of λ_{GSM} is provided in this section.

3.1. General Estimation

The speaker model employed is Gaussian mixture model [4]. Let the parameters for λ_{GSM} be:

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

in which μ_k^{GSM}, Σ_k^{GSM} are mean vector and covariance matrix of the *k*-th Gaussian density function respectively, c_k^{GSM} is the corresponding weight, *M* is the number of mixture components.

Assume there are *N* users currently, whose training data is represented as

$$Y_i = \{ y_1^{(i)}, y_2^{(i)}, \cdots, y_k^{(i)}, \cdots, y_{T(i)}^{(i)} \} \quad i = 1, 2, \cdots, N$$

after being transformed to feature vectors, in which *i* denotes the *i*-th speaker and $T(i)$ denotes the total number of feature vectors. The training data for the $(N+1)-th$ speaker is

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

Estimation of λ_{GSM} is an iterative procedure starting from the initial values obtained by Segmental K-Means Procedure [5][6]. The re-estimation formulas for λ_{GSM} are as follows.

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

(4)

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

(5)

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

(6)

$$\beta_t^{(n)} = \begin{cases} p[y_{t+1}^{(n)}]\, \beta_{(t+1)}^{n} & t = 1 \sim (T(n)-1) \\ 1 & t = T(n) \end{cases} \qquad (7)$$

$$\hat{\mu}_j^{GSM} = \frac{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t) y_t^{(n)}}{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t)} \qquad j = 1,2,\cdots,M \qquad (8)$$

$$\hat{\Sigma}_j^{GSM} = \frac{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t) \cdot (y_t^{(n)} - \hat{\mu}_j^{GSM})(y_t^{(n)} - \hat{\mu}_j^{GSM})^T}{\sum_{n=1}^{N+1}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t)}$$

$$j = 1,2,\cdots,M$$

(9)

Since λ_{GSM} is obtained by using all of the training data of current users, the training time is long and sometimes can not meet real-time needs, especially when system has a large number of users. Thus the real-time estimation is provided in the following sub-section.

3.2. Real-time Estimation

When a new user comes, the training time is consumed mainly on re-training λ_{GSM}. Therefore accelerating the training procedure for λ_{GSM} is important. The real-time estimation updates λ_{GSM} parameters in one step. The initial values are same as those corresponding values of the last new user. The new re-estimation formulas for λ_{GSM} are as follows.

$$\hat{c}_j^{GSM} = \frac{(1-\rho)\sum_{n=1}^{N}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t) + \rho \cdot \sum_{t=1}^{T(N+1)} \theta_j^{(N+1)}(t)}{(1-\rho)\sum_{n=1}^{N}\sum_{t=1}^{T(n)} \alpha_t^{(n)} \cdot \beta_t^{(n)} + \rho \cdot \sum_{t=1}^{T(N+1)} \alpha_t^{(N+1)} \cdot \beta_t^{(N+1)}} \qquad (10)$$

$$j = 1,2,\cdots,M$$

$$\hat{\mu}_j^{GSM} = \frac{(1-\rho) \cdot \sum_{n=1}^{N}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t) y_t^{(n)} + \rho \cdot \sum_{t=1}^{T(N+1)} \theta_j^{(N+1)}(t) y_t^{(N+1)}}{(1-\rho) \cdot \sum_{n=1}^{N}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t) + \rho \cdot \sum_{t=1}^{T(N+1)} \theta_j^{(N+1)}(t)}$$

$$j = 1,2,\cdots,M$$

(11)

$$\hat{\Sigma}_j^{GSM} = \frac{(1-\rho) \cdot A + \rho \cdot B}{(1-\rho) \cdot \sum_{n=1}^{N}\sum_{t=1}^{T(n)} \theta_j^{(n)}(t) + \rho \cdot \sum_{t=1}^{T(N+1)} \theta_j^{(N+1)}(t)} \qquad (12)$$

$$j = 1,2,\cdots,M$$

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

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

$\theta_j^{(n)}(t)$, $\alpha_t^{(n)}$ and $\beta_t^{(n)}$ are computed as (5)-(7). ρ is a weighting coefficient, measuring the contribution of the new registration speech to updating the global speaker model. The greater the value of ρ, the more important the contribution of the new training data. The choice of the value of ρ is important. If ρ is too small, the contribution of new training data will be overwhelmed by the old data and system may be unusable to the new user. On the other hand, if ρ is too large, the system is adjusted to fit the new user, while ignore old users.

Two main points lead to the decrease on computation overhead. In this method, updating is not an iterative procedure. The modification is completed in one phase using formulas (10)-(12). On the other hand updating starts from rather better initial values, avoiding the burden of setting proper initial values, a time-consuming procedure.

4. EXPERIMENTS

4.1. Database and Settings

Data used comes from a Mandarin speech database *863Bag* provided by the *State Education Commission.* Speech data of 50 persons (25 females and 25 males) is used. Each person uttered 50 sentences, ranging from 5.6 seconds to 1.2 seconds.

There are 30 reference speakers (15 females and 15 males). Training data includes 15 sentences. Test data is one sentence. Tests on data of reference speakers constitute closed set test. Tests on data of 20 other speakers who are regarded as outside impostors constitute open set test. The average duration of training data is 60 seconds, and the average duration of every test utterance is 3.5 seconds.

The feature used is 16 cepstrum, 16 delta cepstrum, and delta energy. The general estimation (formulas (4)-(9)) of the global speaker model is adopted in the following experiments.

Equal error rate is used to measure the performance of speaker verification for different scoring methods. The equal error rate is a *posterior* error rate, and at this equal error rate, the decision boundary is set to make the error rate of false rejection be equal to that of false acceptance. 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.

4.2. Statistical Analysis

In this experiment, the likelihood scores of closed set test for GSMSV and CSV are recorded and analyzed. The corresponding statistical results are listed in Table 2.

From table 2, the following conclusions can be obtained:

1) The variance of GSMSV likelihood scores is much smaller than that of CSV method, making the distribution of likelihood scores more compact.

2) The difference between the likelihood scores of GSMSV valid users and impostors is greater than that of CSV method, enlarging the distance between valid users and impostors. Therefore it is more convenient for GSMSV to set a proper threshold.

	Valid speakers		Impostors		Difference
Method	Mean (M_v)	Variance (V_v)	Mean (M_i)	Variance (V_i)	(M_v-V_v)-(M_i+V_i)
CSV	3354.79	584.36	1762.85	730.89	276.69
GSMSV	285.32	188.13	-1312.71	641.47	768.43

Table 2: Statistical analysis of likelihood scores.

4.3. Performance Comparison of Different Methods

In this experiment, the performance of CSV, ASMSV and GSMSV are compared. ASMSV method has the lowest equal error rates when the anti-speaker model consists of all of other reference speakers[3], thus in this experiment set $L = 29$. Table 3 lists the equal error rates.

The equal error rates of both ASMSV ($L = 29$) and GSMSV methods for closed set test and open set test are all significantly smaller than those of CSV method. This demonstrates 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 than that of ASMSV. However the fact that the results of case $L = 29$ are the best results that ASMSV can reach must not be neglected. Experiments in [3] show that when system includes 20 reference speakers and an anti-speaker model consists of 8 and 19 speaker models respectively, the equal error rates for closed set test are 6.44% and 3.65%, and 8.46% and 8.22% for open set test. Therefore when L is smaller, the superiority of GSMSV over ASMSV will be much more prominent.

The average time for verifying an utterance is also listed. The computer used is P-Ⅱ 233. It costs ASMSV over 17 seconds to verify an utterance, while GSMSV spends about 1 second to verify an utterance. Although here ASMSV has a lower equal error rate than that of GSMSV, the verification speed is so slow that system may be intolerable and unusable for practical applications. If ASMSV spent 1 second to verify an utterance, which means that $L = 1$, the equal error rates for closed set test and open set test are 7.80% and 2.16% respectively.

Method	Closed set (%)	Open set (%)	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

Table 3: Performance comparison.

5. CONCLUSION

A novel speaker verification method GSMSV is proposed in this paper. Compared to the CSV and ASMSV methods, GSMSV has the following characteristics. 1) The differences between speakers are enlarged. 2) The system ability to distinguish reference speakers is improved. 3) Impostors can be more easily detected. 4) Verification speed is fast. 5) System is adaptable to speaking speed.

As a summary, the equal error rates of GSMSV method are significantly low, especially for open set test, which is vital for practical applications. GSMSV provides a promising way for realizing verification systems.

6. REFERENCES

1. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juamg, “A vector quantization approach to speaker recognition”, *AT&T Tech. Journal*, vol. 66, pp. 14-26, Mar./Apr. 1987.

2. Zhang Yiying, Zhu Xiaoyan and Zhang Bo. A novel speaker verification method. Accepted by the *Journal of Software*, Chinese edition.

3. Chi-Shi Liu, Hsiao-Chuan Wang and Chin-Hui Lee, “Speaker verification using normalized log-likelihood score”, *IEEE Trans. on Speech and Audio Processing*, 4(1), pp. 57-60, Jan. 1996.

4. Belle L. Tseng, Frank K. Soong and Aaron E. Rosenberg, “Continuous probabilistic acoustic map for speaker recognition”, *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 141-164, 1992.

5. Juang, B. H, *et al*, “Recent developments in the application of hidden Markov models to speaker-independent isolated work recognition”, *Proceedings of IEEE International Conference on Acoustics, speech and Signal Processing*, vol. 1, pp. 9-12, Apr. 1985.

6. L. R. Rabiner *et al*, “Recognition of isolated digits using hidden Markov models with continuous mixture densities”, *AT&T Tech J.* , Vol. 64, No. 6, pp. 1211~1222, July-Aug 1986.