Information Technology Journal 6 (2): 242-246, 2007
ISSN 1812-5638

Counter-Imitation the Speaker's Recognition and the Reconfirmation System Based on GMM

Zhou Ping and Xinxing Jing
Guilin University of Electronic Technology, Guilin, 541004, People's Republic of China

Abstract: It is the foundation of network and information system to strengthen the security for improving status authentication. The veins of voice recognition has merit of no memory, no lose and easy to operate and so on. It can reconfirm the status of the speaker based on counter-imitation recognition and reconfirmation the system so that it can safeguard the security of information and the order. Counter-imitation the speaker's reconfirmation system take the policy-making way which identified the speaker applies to the recognition of confirmation stage. The experimental result had proved the validity and the usability of counter-imitates deliberatively the speaker to reconfirm method.

Key words: Speaker's reconfirmation system, counter-imitation, GMM, reconfirmation system, rejection rate

INTRODUCTION

Along with science technology is developed rapidly and the information technology is used extensively, the foundational and the overall function of network and the information system strengthens day by day, the people enhance a new level in the information security, it is necessary to safeguard the information security. But embezzling the credit number of others, the event of illegal registers network appear continuously. So it is very necessary to strengthen the security for improving status authentication. At present it has the many kinds of methods for the status authentication, the commonly used technical methods include: IC card, password and characteristic of biology recognition, including fingerprint recognition, iris recognition, faces recognition as well as veins of voice recognition and so on. The vein of voice recognition is the most important recognition in the characteristic of biology recognition technology and also is the important side in the speech recognition research field, which pays attention. Compared with other technologies, the veins of voice recognition has merit of no memory, no lose and easy to operate and so on.

Although human's pronunciation has the independence, but it can be imitate. The imitator imitates deliberately speaker's sound when the imitation is similar or same nearly; current confirmation system of pronunciation was deceived by the imitator. At present the main pronunciation imitation methods include: First, imitation is initiative for natural person. For example the special training performers imitate another speaker's tone and tunes are so lifelike on the television frequently. It is very difficult for us to distinguish. Second, the machine imitates. Find out the speaker's the characteristic of the veins of voice, synthesizing pronunciation by using the machine (particular for computer) so that the imitation is completely. Third, the machine and people combine to imitate. Make it sound like another person's pronunciation. This study content: Through comparing the principle and its characteristic of several main imitation methods, proposing the optimized and the fast algorithm in basis of recognition technology of existing speaker and establish special system model about the counter-imitation recognition system of the pronunciation imitation deliberately.

The appearance of the pronunciation imitation threaten information security currently, it is necessary to strengthen security of recognition system of the speaker, it means to research the counter - imitation technology. If it has the counter-imitation technology, the recognition system is reconfirming to speaker's status, so that increasing imitative difficulty to malicious imitator, so it makes the security of information and order better. It is clearly that the counter-imitation technology is very effective in safeguards the information security. As the information technology developing, the social management will be the electron and the automation, the recognition system of speaker which has the counter-imitation technology will get more applications in the important departments and the organizations which are secret.

No matter the recognition system of speaker is based on VQ or GMM (Douglas, 2000), the recognition system of speaker is better than the confirmation system of

Corresponding Author: Zhou Ping, Guilin University of Electronic Technology, Guilin, 541004, People's Republic of China

speaker in performance. The recognition of speaker and the confirmation of the speaker are almost the same in training process, only the method of recognition is different, but two performances is actually bigger difference, the error rate of the confirming system of the speaker is higher than the error rate of the recognition system of the speaker. It shows that the identification system descend slowly when the speaker increases. But according to experiments in the current literature has been already proved. The error rate of the system could get to 1.0% when the population of the recognition system gets to 630 people. But according to experiment has been already proved currently. The error rate of the system could get to 1.0% when the population of the identification system gets to 630 people. From this we can know that the performance of recognition system of the speaker is very well, but the speaker the performance of confirmation system of the speaker is still bad.

This study deals with the reconfirmation system of the speaker in counter -imitates by using above conclusion, because of the recognition system of speaker is better than the confirmation system of speaker in performance, we can use the ways of the recognition system of speaker to advance the performance of the confirmation system of speaker.

The traditional confirmation system of speaker has two shortcomings as follows:

First: The traditional confirmation method of the confirmation system of speaker is to obtain score points of pronunciation from reference the speaker model and then take the scores points compare with one threshold value, so confirm this pronunciation whether is the referenced speaker saying (Nealand *et al*., 2002). Usually, after the experiment the threshold value of being hypothesized in experiment, generally according to the result of the tests hypothesize the threshold value so that the rate of error rejection and the rate of the error acceptance is equal. In order to obtain the appropriate threshold value, the massive experimental work is needed; however hypothesis of the threshold value after the experiment is impossible in actual application and to reduce the adaptability of the system because uses same judge threshold value to the different input pronunciation.

Second: In the practical application, a confirmation system of the speaker is open, some premeditated pronunciation imitators can entry system illegally, then will confirm pronunciation is not any of the referenced speakers to say, this time there is a great change in the judgment effect of the confirmation system, the result is so hard to anticipate.

Based on above two, it has improved for the reconfirm system of the speaker in counter -imitates. The system will not use judgment way of the threshold value and the decision-making way which the speaker identifies applies to the recognition in the confirmation stage, for example: to judge X whether the reference speaker said that, first of all judgment who say in the reference speaker, carries out identifies the process. If the words are said by the referenced speaker, so confirm X to be the alleged referenced speaker *i* to say, or to refuse. It is based on the model of the current speaker. Meanwhile add an overall model of the speaker so that it can reject limitation deliberately by the failure.

ANALYSIS OF THEORY

Considering N reference speakers, their speaker model respectively is λ_1, λ_2, λ_3..., λ_1 obtains scores points by the biggest likelihood $P(O_i|\lambda_i)$, O_i is training data of i reference speaker. When judges section of pronunciations whether the reference speaker I say that, the confirmation strategy is:

$P(X_i|\lambda_i) > Y$ X is said that the speaker I
$P(X_i|\lambda_i) \leq Y$ X is not said that the speaker

Among of them, Y is the decision threshold.
Identifies for the speaker, its strategy is: if

$$i = \text{Argmax } \{P(X|\lambda_K)\} \ (i \leq j \leq N) \quad (1)$$

Then X is said that by I speaker.

But the reconfirmation the system of the speaker in counter -imitates deliberatively applies the identification way of the speaker to confirmation system of the speaker, its judgment method of the confirmation stage is: if

$$j = \text{Argmax } \{P(X|\lambda_K)\} \ (1 \leq K \leq N) \quad (2)$$

When $i = j$, X is said by speaker of number i. When $i \neq j$, X is not said by speaker of number i.

Thus it can be seen, the performance of the confirmation system of the speaker depends on corresponding the performance of the identification system of the speaker in this way. At the same time it has avoided the establish judgment threshold value and then to improve the adaptability of the system. But the system is open, it is possible that non-referenced speaker entry system illegally and imitate deliberately, this time the result of judgment may be wrong, because the pronunciation of confirmation is not said by any of known reference speaker, it is necessary to improve the system.

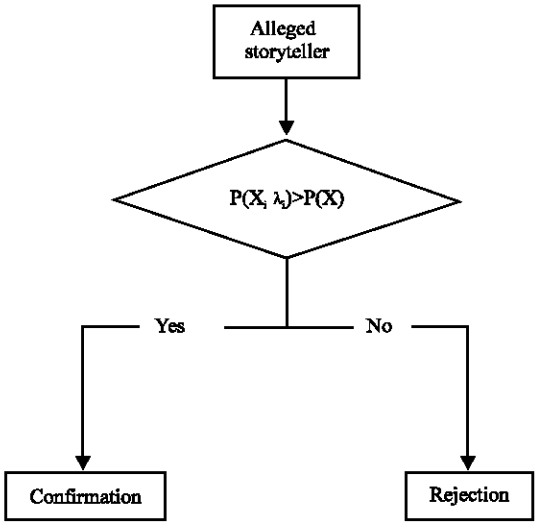

Fig. 1: Improvement distinction strategy

So it has attached an overall model of speaker λ_{N+1} to refuse faitour which in N models of speaker. The parameter of the model λ_{N+1} obtains product

$$\prod_{i=1}^{N} P(O_i|\lambda_{N+1})$$

of scores points by the biggest likelihood; it has represented common characteristic of many speakers (Zhiyou Ma and Yingebun Yang, 2003).

If the pronunciation of confirmation is said by any of reference speaker, that is:

$$P(X_i|\lambda_i)>P(X_i|\lambda_{N+1}) \quad (3)$$

Because λ_i obtains by the speaker's data training, it can be more precise than λ_{N+1} on describing the pronunciation of speaker I in the distribution of acoustics space. But if the pronunciation of confirmation is not any of referenced speakers to say as well as an imitator deliberately, then because of λ_{N+1} has represented many speakers' common characteristic, it is universality, that is:

$$P(X_i|\lambda_i)<P(X_i|\lambda_{N+1}),\ i = 1,2,....N \quad (4)$$

Thereupon improving distinction way as follows: (Fig. 1)

$P(X_i|\lambda_i)>P(X|\lambda_{N+1})$ X is said by speaker of number i, $P(X_i|\lambda_i)\leq P(X|\lambda_{N+1})$, X is not said by speaker of number i.

ESTABLISHMENTS OF MODEL

Training stage:

- Establishes the models λ_i for each speaker, the training criterion is the biggest likelihood can get points $P(O_i|\lambda_i)$, among of them O_i is all training data of the I the referenced speaker, for i= 1, 2... N;
- Establishes a model of overall speaker λ_{N+1}, the training criterion is the biggest likelihood can get point

$$\prod_{i=1}^{N} P(O_i|\lambda_{N+1})$$

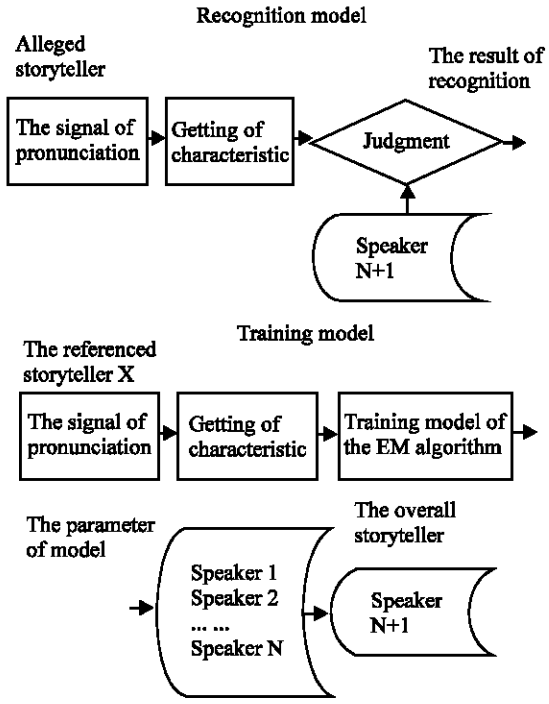

Fig. 2: Counter-imitates deliberately the speaker to reconfirms system structure drawing

The stage of Confirmation

- Calculates $P(X_i|\lambda_i)$ and $P(X|\lambda_{N+1})$
- If $P(X_i|\lambda_i)>P(X|\lambda_{N+1})$, then confirmed that X is the pronunciation of referenced speaker i, or to refuse X is the pronunciation of the referenced speaker. The speaker of counter -imitates deliberately to confirm the system as shown in Fig. 2:

Model of the reconfirmation system: The reconfirmation system of the speaker of the counter -imitates deliberately is used by Gauss mixed model to experiment, it means each referenced speaker's training pronunciation is expressed by Gauss mixed density in the distributions of acoustics space. The parameter of each speaker's model as follows:

$$\lambda = \{w_i, \mu_i, \Sigma_i\},\ i = 1,...,M$$

Among of them, w_i, μ_i express separately the k average value and covariance matrix of Gauss function

vector, Σ_i is the k right of Gauss function, M expresses how many mixed Gauss function.

If $O = \{o_1, o_1, \ldots, o_T\}$ then

$$b_i(X) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)^t \Sigma_i^{-1}(X-\mu_i)\right\} \quad (5)$$

$$P(O|\lambda) = \prod_{k=1}^{D} P(o_k|\lambda) \quad (6)$$

Among of them, D is dimension of the eigenvector.

METHODS OF EXPERIMENTAL

Establishment of the pronunciation imitation database: In order to confirmation the system of counter -imitates deliberately, it is necessary to build a pronunciation imitation database. In practical application, because the confirmation system usually opens and then the premeditated pronunciation imitator (other speakers besides this speaker) includes:

- The speaker of unknown pronunciation out of the set (refers to unknown illegal users).
- The pronunciation of speaker was recorded in advance out of the set (experimental, record voluntarily).
- The other speakers in set.

When system is operating, it has not the material of the speaker of unknown pronunciation out of the set. So the pronunciation imitation database consists of set 2 and 3. There are 60 men in experimental storehouse of this article, divide into two sets: A (marking is 1-30) and B (marking is 31-60), sees the next table. It has 30 people in each set, each people will record ten digital strings and a character string (all are standard Chinese, the content are the same). Among of them, the marking 1 speaker in set A is the man who will be imitated, 29-30 are constituted by the free speakers; 30 people (marking 31-60) in the set B is requested to imitate speaker 1 intentionally, their ages, their education level of these 30 people are similar as the speaker 1, the native place is the same. It is mainly too sure to improve the similarity of the speaker 1 in the set B so that it can increase the difficulty of recognition.

Table 1: Experiment uses the database

Pronunciation imitation storehouse	Set A	Set B
Record person	1, 2 - 30	31-60
Training data	10 numerals strings, 1 character string	
Test data	10 numerals strings, 1 character string	

Method of the experimental: First, testing (Closed set Test) set A in the experiment, then testing (Open Set Test) with set B to set A (Table 1). Error rate takes average value of error rate of the reference speaker. The recording environment is the quiet room, each person's training sentence is equal to 20 s and each numeral string of test is equal to 1s. All pronunciations are sampled of the 8 KHZ, uses the filter $H(Z) = 1\text{-}az^{-1}$ ($a = 0.95$) to pre-aggravate the digitized pronunciation. 1 frames each 32 ms, the frame moves is 16 ms and each signal add the window in hamming window. The eigenvector is 16 steps LPC reversal Spectrum and MFCC, respectively. The model of speaker is the mixed the Gauss model of the 64 Gauss mixed density; the training method uses the standard the EM algorithm.

EXPERIMENTAL RESULTS

Test A: In this experiment, we choose set A as the closed set and set B as the open set. When the open set is testing, all the 30 people (marking 31-60) are Alleged speaker 1 in B set, the data of test and training are the same. The result of the confirmation method of the speaker which based on traditional GMM and the confirmation method of the speaker which based on counter -imitates deliberately show in Table 2.

From the Table 2, the acceptance rate of its error is much lower than the acceptance rate of traditional GMM method error, whether test in the closed set or the open set based on counter-imitation method of dynamic threshold value. Especially to closed test, when the training sentences are shorter, the acceptance rate of its error has got to 0.0% and the rate of its error rejection is much lower than the error rejection rate of traditional method in the same error threshold value. It means that it has very strong ability of distinguish speakers by the counter - imitation system.

Test B: It has proved that the validity of the new method through experiment (Zhen and Wei, 2002). In order to prove the superiority of this method compared to the traditional method, this article has carried on the following experiment: through modifying the threshold value of the traditional method which based on GMM, make its rejection rate and acceptance rate of error get to the corresponding number of the counter - imitation method.

Table 2: Tests A result of the experiment

The method of confirmation	GMM method (%)		Counter-imitation method (%)	
Test	FR	FA	FR	FA
The closed set	12.1	12.1	5.9	0
The open set	--	21	--	0.01

Table 3: Tests B result of the experiment

The method of confirmation test	GMM method (%)		GMM method (%)		Counter-imitation method (%)	
	FR	FA	FR	FA	FR	FA
The closed set	5.9	19.7	62.2	1	5.9	0
The open set	--	28.1	--	5	---	0.01

Now comparing with corresponding value of the traditional method in error acceptance rate and the error rejection rate, the result of experiment are shown Table 3.

From the result of experiment shows that the traditional the GMM method achieved respectively corresponding value of the counter- imitation method on the rejection rate of error and the acceptance rate of error, it has increased greatly on the acceptance rate of error and the rejection rate of error; it is higher than the corresponding acceptance rate of error and the rejection rate of error of the counter- imitation method.

THE EXPERIMENT ANALYZES AND CONCLUSION

The result of experiment had proved validity and usability of reconfirmation method of the speaker's counter-imitates deliberatively. This method will reduce the acceptance rate of error to minimum and reduce greatly the rejection rate of error, enhance versatility of the system. The traditional confirmation system of the speaker is strict in the secret, when the acceptance rate of error is reduced, the rejection rate of acceptance will be raised. Because it is limited by the threshold value, the rejection rate of error and the acceptance rate of error affect and restrict each other; one reduced causing the other one raised inevitably. But this article uses the method has overcome limitation of the traditional method well. While the acceptance rate of error is reducing, it hasn't caused the rise of the rejection rate of error.

REFERENCES

Douglas, A.R., 2000. Speaker verification using adapted gaussian mixed models, Digital Signal Processing, 10: 19-41.

Nealand, J.H., A.B. Bradley and M. Lech, 2002. Discriminative feature extraction applied to speaker identification. Signal Processing, 2002 6th International Conference on, 1: 26-30.

Zhiyou Ma and Yingebun Yang, 2003. Further Feature Extraction for Speaker Recognition. Systems, Man and Cybermetics. IEEE International Conference on, Vol. 5, Oct. 5-8.

Zhen, Y. and L.C. Wei, 2002. A New feature extraction based the reliability of speech in speaker recognition, Signal Processing, 2002 6th International Conference on, 1: 26-30.