A TEXT-PROMPTED DISTRIBUTED SPEAKER VERIFICATION SYSTEM IMPLEMENTED ON A CELLULAR PHONE AND A MOBILE TERMINAL

Tsuneo Kato and Hisashi Kawai

KDDI R&D Laboratories Inc.
2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502, Japan
e-mail: tkato@kddilabs.jp

10.21437/Interspeech.2006-177

Abstract

For a practical application of biometrics authentication on cellular phones, a speaker verification system was implemented on a cellular phone and a PDA type mobile terminal. The system has the following three features: 1) a distributed system configuration for connectivity to internet services, 2) text-prompted speaker verification using connected digit patterns for robustness to imposture, and 3) incremental model update for preventing deterioration of accuracy. The system was examined in a two-month test, and effects of the incremental model update were evaluated with those data. Experimental results showed that the average equal error rate over this period was reduced from 8.6% to 4.4% by monthly update, to 2.9% by weekly update, and to 0.9% by daily update.

Index Terms: speaker verification, implementation, model update

1. Introduction

Speaker verification is a promising biometrics authentication for cellular phones, because the current text input system with numeric keys takes time, and speech input from a regularly equipped microphone is expected to be more popular and commonly used. To evaluate the performance and extract issues on a practical application, a speaker verification system was implemented on a cellular phone after an implementation on a PDA type mobile terminal. Speaker verification on cellular phones have some problems to be solved, such as 1) difficulty in providing authentication function for internet contents, 2) password utterances unwanted in public, and 3) an accuracy limitation and deterioration over time.

In terms of the first problem, a cellular phone system using a conventional telephony speaker verification server makes users wait for tens of seconds by switching between packet communication and circuit switching, and it is not practical. A distributed speech recognition (DSR) [1] system configuration, in which acoustic features are transmitted to a server and processed there, becomes a solution to this problem. To the second problem, the text to be pronounced has to be changed to meaningless words because users are afraid that their passwords are intercepted. Instead of the password type verification, text-prompted speaker verification [2], in which the system specifies a changing text for verification, has been proposed to prevent imposture by playing recorded client's speech. We have proposed an algorithm for creating effective connected digit patterns for text-prompted speaker verification [3]. Regarding to the third problem, natural change of human voice causes the deterioration of accuracy. Many kinds of effective speaker adaptation techniques [4, 5] have been proposed to produce highly discriminating client models with limited adaptation data. To compensate the deterioration, incremental model update with verification utterances using the speaker adaptation techniques has also been examined [6]. Since the incremental model update is not examined on text-prompted speaker verification using connected digit patterns, the effectiveness is expected to be evaluated on the system where digits are re-ordered in verification.

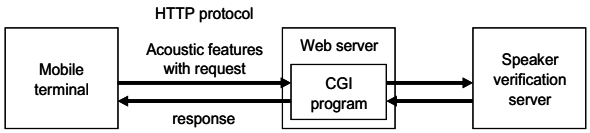

Figure1 *Distributed system configuration.*

To the problems in practical use of speaker verification systems for cellular phones, we have implemented all the above-described techniques on our prototype system on the PDA type mobile terminal and the cellular phone. A standard acoustic feature extraction function is implemented on the terminals with speeding-up techniques. The effectiveness of these techniques was evaluated in a two-month test in a normal office environment.

In section 2, the system configuration and features are overviewed. The hardware and the user application are described. In section 3, the incremental model update with verification utterances on this system is explained. In section 4, the performance on the PDA type mobile terminal is evaluated with focusing on the incremental model update.

2. System overview

2.1. System configuration and features

2.1.1. Distributed system configuration

Configuration of the authentication system is a distributed (client-server) type shown in Figure 1. Acoustic feature extraction from the microphone input is processed on the terminal, then the acoustic features are transmitted to the speaker verification server via a web server, and enrollment or verification is processed on the speaker verification server. The results of enrollment and verification are returned via the web server. This system configuration was designed for connectivity to web-based applications or web contents. The authentication step can be easily attached to existing web-based applications or web contents by a hyperlink to the CGI program which issues requests to the speaker verification server. The acoustic features transmitted from the mobile terminal to the web server in HTTP protocol are DSR acoustic features standardized in ES201108 [7] of ETSI.

2.1.2. Text-prompted speaker verification using connected digit patterns

In the text-prompted speaker verification, input speech is first processed by speech recognition. If the recognition result corresponds with the prompted text, it proceeds to enrollment or verification. If the recognition result is different from the prompted text, the enrollment or verification is terminated as failure. Text-prompted speaker verification is more robust to imposture by playing recorded client's speech because the verification pattern changes every time.

This system prompts a 6-digit pattern or an 8-digit pattern for enrollment and 4-digit patterns for verification. The six or eight digits of the enrollment pattern are phoneme-balanced, and the verification patterns preserve partial digit sequences of the enrollment pattern to improve the accuracy [3]. By use of connected digit patterns, utterances for enrollment are kept only five since all the verification patterns are combinations of ten digits, and there is no need to collect various sounds. Users' hesitation for utterance is less than the case of the password type speaker verification because the specified text is a meaningless connected digit pattern.

Moreover, a decision procedure using multiple patterns [3] was implemented to the system for reducing incorrect decisions. Commonly, the result of verification is returned "accept" or "reject" with an utterance. However it is sometimes difficult to decide "accept" or "reject" with an utterance. Therefore the server set three status, "accept", "reject", and "not decided". When the decision is difficult with an utterance and "not decided", the system requests an additional utterance with a different connected digit pattern to make a reliable decision.

2.1.3. Incremental model update

To improve the verification accuracy, incremental model update using verification utterances is employed. The verification accuracy deteriorates not only by change of physical conditions but also by natural change of voice, which is called aging of speech. To prevent this deterioration, client models have to be updated. Compulsory periodical re-enrollment procedure is painful to users and not practical. Therefore incremental model update using verification utterance is examined instead of periodical re-enrollment. The method is detailed in section 3.

2.2. Authentication applications on mobile terminals

The user application was implemented first on 1) a PDA type mobile terminal based on Ubiquitous Communicator (UC) [8] of YRP Ubiquitous Networking Laboratory (YRP-UNL), and then to 2) a CDMA2000-1X EVDO cellular phone.

On the PDA type mobile terminal CDMA2000-1X EVDO card was newly added. Its CPU is Hitachi SH-3 144MHz and the memory is 32MB. The cellular phone is based on ARM 9 processor. To realize a quick response, the ETSI201108 acoustic feature extraction program is implemented with speeding-up techniques, such as fixed point calculation, architecture dependent efficient instructions. Reducing the processing time less than the utterance time, the acoustic features are sent to the server immediately after an utterance ends. The response from the server returned within 5 seconds.

The GUIs of the user applications on the PDA type mobile terminal and the cellular phone are shown in Figure 2.

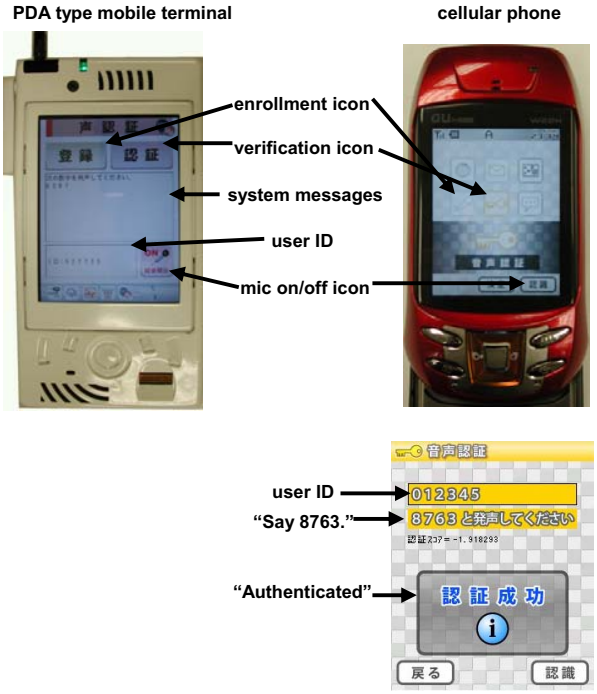

Figure 2 *GUIs of the authentication applications.*

Users start the enrollment procedure by selecting the "enrollment" icon, and start the verification from the "verification" icon. A user ID and messages from the system are displayed. The icon at the lower-right corner is an ON/OFF icon of the microphone.

In the enrollment procedure, a 6-digit pattern or an 8-digit pattern is specified and five utterances are requested to the user. The server recognizes the user's utterances and enrolls acoustic features of the speech as speaker-dependent digit HMMs. The result of enrollment is displayed on the terminal. In the verification procedure, the system prompts a 4-digit pattern which changes every time. The server first recognizes what the user has said. If the recognition result is different from the prompted pattern, the utterance is rejected. If it corresponds with the prompted pattern, the server computes a likelihood score using the speaker-dependent digit HMMs of the user ID. One of the three status, "accept" or "reject", or "not decided" is determined by the score. If an utterance is "not decided", an additional utterance is requested.

3. Incremental model update

3.1. Basic enrollment and verification process on the server

The speaker verification server executes enrollment and verification according to the requests from the terminals.

In enrollment, four-mixture continuous-density speaker-dependent digit HMMs are trained based on speaker-independent digit HMMs. The mixture weights and mean vectors of Gaussian distributions are re-estimated by ML-based training.

In verification, a decision is made based on a normalized log likelihood score *S*, which is a time-normalized difference

between two log likelihood scores of the client model (speaker-dependent HMMs) and a world model (speaker-independent HMMs). For the decision procedure using multiple patterns, the server sets two thresholds for S to decide one of the three status, "accept" or "reject" or "not decided". If an utterance is "not decided", an additional utterance is requested to the user to make a reliable decision.

3.2. Incremental update of client models

Incremental model update is executed by adding a verification utterance to the list of enrollment data, re-estimating the parameters of client models and updating them. Since the enrollment data increase gradually over time, maintenance of the initial accuracy, or rather improvement is expected. For preventing misguided adaptation to impostors' speech, only the verification utterances certified with high certainty are added to the list. A threshold S_u is set for determining whether the utterance is to be added to the list or not. If the recognition result of a verification utterance is different from the prompted pattern, or the score S is lower than the threshold S_u, the utterance is not added to the list and the client model is not updated.

An essential parameter for the effects of incremental model update is frequency of the model update. The performance is to be evaluated as a function of update frequency.

4. Evaluation of the system

4.1. Test conditions

The speaker verification system was tested with the PDA type mobile terminal by 15 male users for two months. The users enrolled their voice once a day and verified their voice five times, except for weekends and holidays. All the speech data for enrollment and verification were stored in the server to reuse in offline simulations.

The enrollment pattern specified by the system was a fixed 6-digit pattern and common to all the users. The verification patterns consist of four digits out of the six of the enrollment pattern. Because the verification patterns were common to all the users, the utterances for verification were used as imposture data for other users in offline simulation.

In the following evaluation, equal error rates (EERs) were calculated without regarding to the decision procedure using multiple patterns.

4.2. Evaluation of accuracy without incremental model update

Figure 3 shows false acceptance rate (FAR) and false rejection rates (FRRs) as a function of a threshold determining "accept" or "reject" without incremental model update. FRRs are plotted with different intervals between enrollment and verification, because FRR is dependent on the interval, while FAR is independent. The error rates are the mean values of the 15 users' utterances.

FRR on the day of enrollment is significantly lower than that of other days. The reason is thought that utterances for enrollment and those for verification on the same day are acoustically quite similar. From after 3 days to after 30 days, FRRs are similar, but after 55 days, FRR increases. The increase of FRR is caused by decrease of the clients' score S.

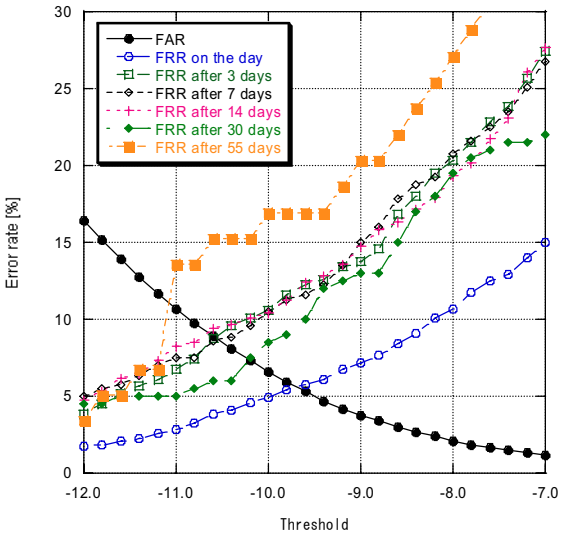

Figure 3 *FRR increasing with an interval between enrollment and verification.*

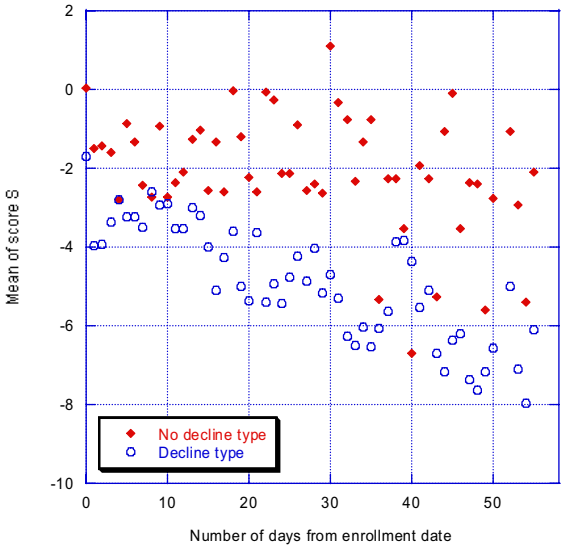

Figure 4 *Two types of a client's score* S *with relation to aging. No decline type and decline type.*

Investigating the score S for individual clients, the declining trend over days greatly depends on individual clients. The individual trends are classified into two types roughly. Figure 4 shows the examples of the two types. Mean of score S of "No decline type" fluctuates but has no clear declining trend, while the other has a clear declining trend. Ten users are classified in "No decline type", whereas five in "Decline type".

4.3. Evaluation of accuracy with incremental model update

The incremental model update was applied to the system for preventing the declining trend of the clients' score S shown in Figure 4. The following five update frequencies were evaluated.

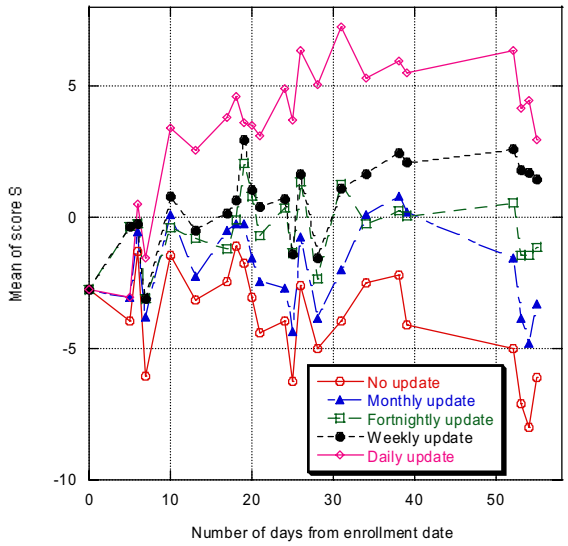

Figure 5 *Improvement of the mean score* S *of client utterances by the incremental model update.*

Table 1. *Average equal error rates (EER) of the period with different update frequencies.*

Update frequency	Equal Error Rate (EER)
No update	8.6%
Monthly update	4.4%
Fortnightly update	4.1%
Weekly update	2.9%
Daily update	0.9%

a) No update
b) Monthly update (every 30 days)
c) Fortnightly update (every 14 days)
d) Weekly update (every 7 days)
e) Daily update (every day)

At certain update frequencies, one of the five utterances for verification is selected randomly and verified if it is to be added to the list of enrollment data. The threshold S_u was set at -9.0 so that the FAR is under 4% from Figure 3.

Figure 5 shows trajectories of the mean score S for the five update frequency cases. The S values are the mean of all the clients. The mean score of "No update" decreases from the enrollment day. In contrast, the mean scores increases in the update cases. The increase is the most in "Daily update" case because the training data are collected at the fastest rate. Even in the case of "Fortnightly update", the score increases from the enrollment day.

Finally, equal error rates (EER) in the five update frequencies are compared. Table 1 shows average equal error rates (EER) of the period. The EER of "Weekly update" is calculated by the average of from 1-day interval to 7-days interval. In a similar way, the EER of "Daily update" is that of 1-day interval. The average EER is greatly reduced by incremental model update, from 8.6% to 0.9% by "Daily update" and to 2.9% by "Weekly update". Comparing the EERs of incremental model update cases with the EER on the enrollment day (5.5%) in Figure 3, the EERs of update cases are still less than the EER on the day of enrollment. The reason is that the average enrollment data for the update cases are much more than five utterances on the enrollment day.

5. Conclusions

A text-prompted distributed speaker verification system was implemented on a cellular phone after an implementation to a PDA type mobile terminal. The distributed system configuration using HTTP protocol between the terminals and server provided connectivity to web services. A quick response within 5 seconds was realized by an implementation of acoustic feature extraction program with speeding-up techniques. The text-prompted speaker verification using connected digit patterns provided not only robustness to imposture by playing recorded client's speech, but also reduction of hesitation for pronunciation in public. As a measure to deterioration of accuracy, incremental model update using verification utterances was implemented.

Based on the two-month test data of 15 users with the PDA type mobile terminal, change of each client's error rate over time and effects of incremental model update were investigated. With the incremental model update, the mean score S was improved with increasing training data, whereas it declined without update. The average EER over this period was improved from 8.6% to 0.9% by daily update and to 2.9% by weekly update. Because a probability of misguided adaptation to impostors has not been evaluated, simulations of imposture attacks are left as a future work.

6. References

[1] D. Pearce, "Enabling New Speech Driven Services for Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-ends," *Proc. AVIOS 2000*, 2000.

[2] T. Matsui and S. Furui, "Concatenated Phoneme Models for Text-variable Speaker Recognition," *Proc. ICASSP 93,* vol. 2, pp.391-394, 1993.

[3] T. Kato and T. Shimizu, "Improved Speaker Verification over the Cellular Phone Network using Phoneme-balanced and Digit-sequence-preserving Connected Digit Patterns," *Proc. ICASSP 2003*, vol. 1, pp.57-60, 2003.

[4] J. L. Gauvain, C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," *IEEE Trans. on SAP*, Vol. 2(2), pp.291-298, 1994.

[5] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models," *IEEE Trans. on ASSP*, Vol. 39, No. 4, pp.806-814, 1991.

[6] C. Fredouille, J. Mariethoz, C. Jaboulet, J. Hennebert, "Behavior of a Bayesian adaptation method for incremental enrollment in speaker verification," *Proc. ICASSP 2000,* pp.619-623, 2003.

[7] "ETSI ES 201 108 v1.1.2 Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm", 2000.

[8] K. Sakamura and N. Koshizuka, "T-Engine: the open realtime embedded systems platform," *IEEE MICRO,* vol. 22, no. 6, Dec. 2002.