THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.4E 2003. 12. pp. 176~182

Voice Verification System for m-Commerce on CDMA Network

Youn-Jeong Kyung*
*Tokyo Institute of Technology, SK Telecom Platform R&D Center
(Received October 22 2003; accepted December 3 2003)

Abstract

As the needs for wireless Internet service is increasing, the needs for secure m-commerce is also increasing. Conventional security techniques are reinforced by biometric security technique. This paper utilized the voice as biometric security techniques. We developed speaker verification system for m-commerce (mobile commerce) via wireless internet and wireless application protocol (WAP). We named this system the mVprotek. We implemented the system as client-server architecture. The clients are mobile phone simulator and personal digital assistant (PDA). The verification results are obtained by integrating the mVprotek system with SK Telecom's code dimension multiple access (CDMA) system. Utilizing f-ratio weighting and virtual cohort model normalization showed much better performance than conventional background model normalization technique.

Keywords: *Speaker verification, WAP, PDA, m-Commerce, Cohort model*

I. Introduction

Using the internet both as a source of information and as an e-commerce transaction has become astonishingly routine in the past few years since the world wide web has been introduced. People navigate the internet many times a day for various reasons, and the internet usage is increasing exponentially. But accessing the internet through the typical wired routes are pretty cumbersome, especially when we're away from our desktop computers and wire-based networks. What if we could access the Internet through smaller, more ubiquitous and, frequently, wireless devices? The information - weather, stock quotes, news, sports scores, train and airline schedules - would be available anytime and anywhere. Transactions such as making reservations, banking, stock trading and shopping could be done at one's convenience. Due to convenience, the needs for wireless internet service is ever increasing.

Corresponding author: Youn-Jeong Kyung (yjkyung@furui.cs.titech.ac.jp)
Tokyo Institute of Technology, SK Telecom Platform R&D Center
2-12-1 Ookayama Meguro-ku Tokyo 152-8552 Japan

The requirements of security are also increasing in accordance with widespread of wireless internet service such as WAP. Most widely used means of individual verification for e-commerce (or m-commerce) security such as security card, electric signature and passwords have the risk of robbery or forgery. A biometric security system such as fingerprints has the disadvantage of higher cost of additional equipments. But the verification method by one' s speech is very economic and convenient methods for e (or m) commerce. Our work aims to develop an authentication system by using speaker verification technique. We referred this system to the mVprotek[1].

II. The mVprotek System Overview

2.1. Limits & Solutions

To develop the mVprotek we must consider some limits. Most of limits are caused by wireless network circumstance.

- Bandwidth: Clearly there are obstacles to this vision, and one important fact is the limited graphical bandwidth of handheld devices, such as cellular phones and PDAs, compared to desktop and laptop computers. Since the m-commerce use data channel instead of voice channel, we must consider the bandwidth and data transfer rates.

To alleviate this problem we suggest to use embedded enhanced variable rate coder (EVRC). In cellular network of CDMA by SK Telecom, the EVRC is used as vocoder. It can compress the speech signal into 1.2 kbps ~ 8 kbps.

- Security: Since the speaker verification technique is applied to m-commerce. The security performance is very important.

To solve this requirement, we consider the smart confirmation technique - customer interface -. This technique is helpful to improve the verification performance and is patent pending[2]. The technique is explained in detail in chapter 4.

2.2. Network Configuration

The m-commerce environment is CDMA (IS-95B) wireless internet in Korea by SK Telecom. The network configuration presented in Fig. 1.

The client (custom) is cellular phone (include WAP phone) or desktop personal computer (PC). The m-commerce server is accessible via two ways. One is through only internet protocol (IP) hub. In this case, the protocol is transmission control protocol / internet protocol (TCP/IP) and the server is implemented by hypertext markup language (HTML). The other is through WAP gateway. In this case, the protocol is WAP and it must be communicate with client using wireless markup language (WML).

2.3. Speaker Verification Algorithm

- End-point detection: We use general end-point detection algorithm (by zero crossing rate (ZCR) etc.) and additional post-processing algorithm. It depends on formant and pitch. We reject the noise or cough by this algorithm. Also we implemented this algorithm into cellular phone simulation. The cellular phone has lower computing power than desktop computer. This algorithm doesn't need complicated computation. It is also patent pending[3].
- Speaker verification: We make the speaker model using virtual cohort model. Also we use the f-ratio weighting function to improve verification performance. It is presented in detail in Chapter 3.
- Speech recognition: In our confirmation scenario, since the speaker verification system is implemented with text prompted methods. Speech recognition stage is required. We utilized hidden markov model (HMM) based speech recognition module.

2.4. Implementation

The m-commerce server is connected on public internet.

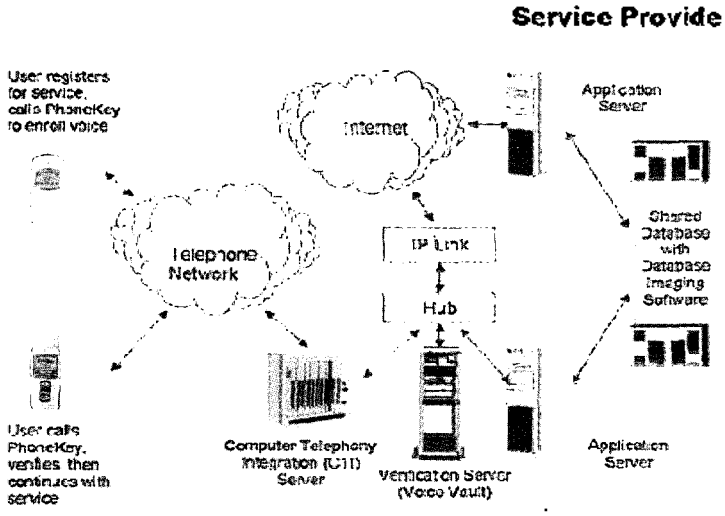

Figure 1. The configuration of SK Telecom CDMA wireless internet.

The clients are implemented for 3 different devices type. One is cellular phone simulation. We programmed cellular phone into laptop computer. The EVRC codec and end-point detection algorithm is also simulated. We utilized the cellular phone as wireless modem. The CDMA (IS-95B) network of SK Telecom is providing '1501' service. Second client is PDA (we use iPAQ by COMPAQ). For implementation, we programmed EVRC codec and end-point detection algorithm using visual studio for WinCE. The centrol processing unit (CPU) of PDA is strong ARM chip and it doesn't have any good quality micro-phone. Third client is WAP phone simulator. We use the Nokia's WAP toolkit version 2.0.

III. Speaker Verification Algorithm

3.1. Virtual Cohort Normalization

Virtual cohort normalization was proposed as a new cohort normalization method for HMM based speaker verification[4]. It has been known that score normalization using the likelihood ratio of the reference speaker model and speaker background model or cohort models is very effective for enhancing the performance. In the conventional score normalization methods, cohort models are determined by choosing the closest speaker model to the reference model among the other speaker models or combining some speaker models closer to the reference model. But these methods have many difficulties to finely control the likelihood variation of cohort models, because constituent unit of cohort set is obliged to be "Speaker model". Therefore, it can be considered that the likelihood score ratio is not stable. We used a new constructing cohort set and the way of synthesizing cohort models which are focusing on the acoustic similarity between models in fine-structure level. Fig. 2 shows conceptual illustration of virtual cohort model construction method used in our mVprotek system. This situation in Fig. 2 shows that three simplified speaker models (A, B, C) are closer to the reference speaker model I. Speaker model V is virtually constructed as cohort model using some of the

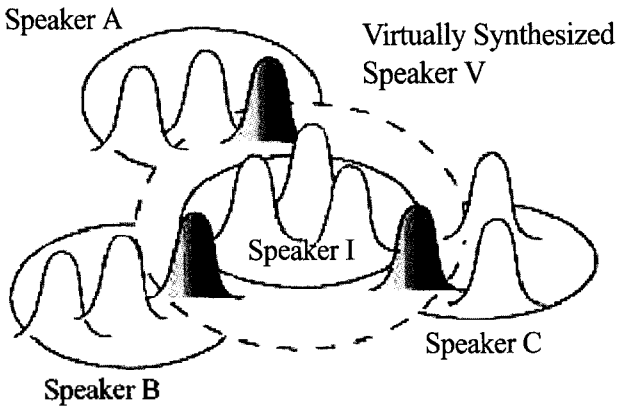

Figure 2. The concept of cohort model construction.

closer model's distributions. In Fig. 2, gaussian pdf shadowed with gray. As shown in Fig. 2, virtually synthesized cohort model V is statistically closer to the reference model than cohort set or cohort models selected by the conventional speaker-based method. These selections are determined by distance between distributions, which mean the similarity of local acoustic features. In mVprotek system, we used the Bhattacharyya distance as a measure of similarity between the distributions.

In this paper, we introduce distribution-based selection, one of virtual cohort model construction methods, used in mVprotek system. A model λ of virtual cohort model is defined as follows.

$$\lambda_p = \{a_{p,s,s}, a_{p,s,s+1}, w_{p,s,m}, N_{p,s,m}\}_{s=1,\cdots S;m=1,\cdots M} \quad (1)$$

where p is phoneme model, $a_{p,s,s}$ is state transition probability from state s to s, $w_{p,s,m}$ is a weighting coefficient of state s and m-th mixture and $N_{p,s,m}$ is a gaussian pdf of state s and m-th mixture.

Therefore, a synthesized k-th cohort speaker model is defined as follows.

$$\lambda^{(c_k'(I))} = \{a_{p,s,s}^{(c_k'(I))}, a_{p,s,s+1}^{(c_k'(I))}, w_{p,s,m}^{(c_k'(I))}, N_{p,s,m}^{(c_k(I,p,s,m))}\}_{p=1,\cdots P;s=1,\cdots S;m=1,\cdots M} \quad (2)$$

where $c_k'(I)$ represent the k-th cohort speaker model virtually synthesized for the reference speaker I and $c_k(I,p,s,m)$ is the k-th cohort speaker.

The transition probability is defined as

$$a'^{(c_k(l))}_{p,s,s} = \frac{\sum_m a^{(c_k(l,p,s,m))}_{p,s,s}}{\sum_{j=0,1}\sum_m a^{(c_k(l,p,s,m))}_{p,s,s+j}} \quad (3)$$

and the weighting coefficient is

$$w'^{(c_k(l))}_{p,s,m} = \frac{w^{(c_k(l,p,s,m))}_{p,s,m}}{\sum_m w^{(c_k(l,p,s,m))}_{p,s,s+j}} \quad (4)$$

The probabilities for self-loop state transition and weighting parameter are re-normalized using Eq. 3 and Eq. 4 according to the constraint given by HMM.

3.2. F-Ratio Weighting

The cepstrum parameter weighted by f-ratio was adopted to maximize a discrimination between voice of speakers. The f-ratio was used as a criterion of weights to characteristic parameter of each user's voice, and it is represented by a ratio of the variance of voice for inner speaker and the variance of voice for intra speaker. See following Eq. 5.

$$F-ratio = \frac{Variance\ of\ Speaker\ Mean}{Mean\ of\ Intraspeaker\ Variance} \quad (5)$$

If a feature parameter has a high variation of voice for intra speaker and a low variation of voice for inter speaker, the feature parameter has a high F-ratio and can be considered that it has an effective verification capability towards voice of speakers.

$$\text{F-ratio} = \frac{Var(E(C_{ij}))_{total-speaker}}{E(Var(C_{ij})_j)_{total-speaker}} \quad (6)$$

F-ratio was applied on a cepstrum segment basis (26 order) as shown Eq. 7.

$$b_j(o_t) = \sum_m^M \frac{w_{jm}}{\sqrt{2\pi\sigma^2}} \exp\frac{((x-\mu)(1/\text{F-ratio}))^2}{2\sigma^2} \quad (7)$$

where $b_j(o_t)$ is a probability of observation o_t vector in the state j, and M is the number of mixture. The f-ratio vector gives weight to the correspondent vector $x-\mu$.

IV. The mVprotek Architecture

The mVprotek server has some functions.

- Enrollment function: For using the speaker verification function the customer must be registered into mVprotek system. The customer chose the 5 private question and reply that question via voice. These questions and answers are stored in customer DB with encapsulation.
- Training function: The mVprotek make the user model and cohort model automatically using customer's utterances.
- Verification function: The mVprotek choose and transferred to the client the question from customer DB randomly. It is like the text prompted style. The utterance is through the speech recognition on stage 1 and then compare with user model in stage 2. In the stage 1, the system check if the answer is correct or not. If the answer is correct, the utterance is compared with user model in stage 2.
- Adaptation function: The off-line re-training function is implemented to guarantee the performance and to give robustness against time varying characteristic of one's voice.
- Management function: This function controls the other functions and manages the customer database (DB).

The mVprotek call flow is showed in Fig. 3.

V. Experimental Results

5.1. The Speech Database

The speech database for test of speaker verification has been recorded by mobile phone via CDMA. It considered speaker's gender, age, recording session, even model of mobile phone and mobile phone carriers. In order to capture the variation in speaker's voices over time, all speakers recorded in 6 or 7 sessions and each session is separated by 2 weeks at most. It is one of the most important aspects of speaker recognition[5]. Total 182 speakers, which is 110 males and 72 females, were asked to pronounce a list of 50 words which consist of 2~4

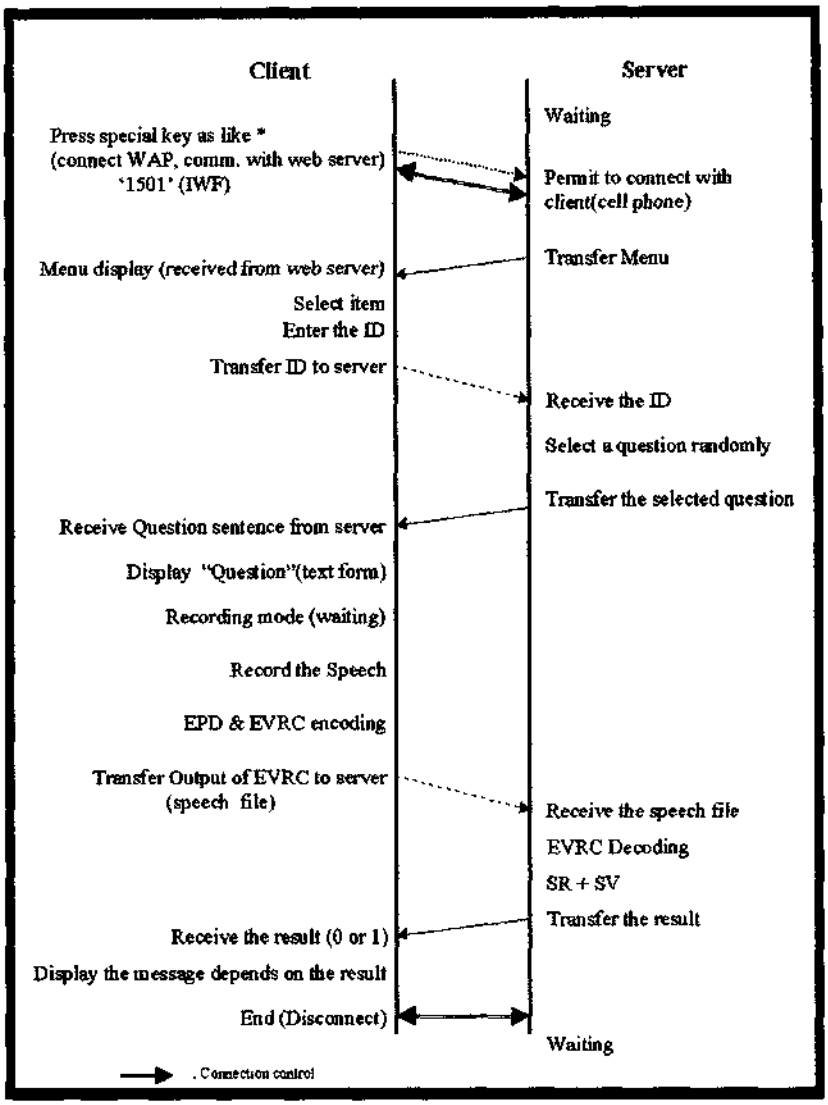

Figure 3. The mVprotek call flow.

syllables and 50 of four connected digit for each session in a quiet environment. Total speaker's speech data were converted to 8 kHz, 16 bit, mono, Intel pulse code modulation (PCM, (LSB,MSB)) format. 33 out of 182 speakers were used as imposters and the other 149 speakers were used as enrolled users. Three sessions were used to train the speaker model and background model. The remaining sessions were used for testing and adaptation.

5.2. Experimental Results

The speech features were extracted on a frame rate of 10 msec and a frame size of 30 msec. The pre-emphasis (0.97) and hamming window were used. This feature vector includes 12 mel-cepstrum coefficients, the energy and first order derivatives.

The mVprotek is the text prompted speaker verification system, used HMM models for speaker verification. The background model and speaker model were trained using the EM algorithm on speech data from mobile phone as described in the previous subsection. All speaker and background models were designed with 38 mono phoneme models. The threshold of verification normalizing score was set to uniform with 0 for all speaker verification.

The first experiments were to compare the performance of virtual cohort normalization using f-ratio with back-

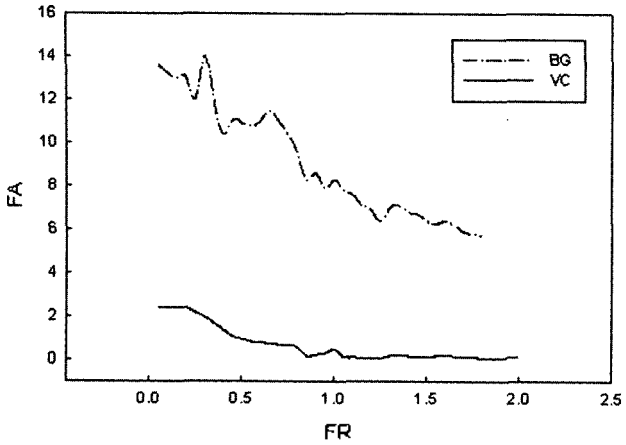

Figure 4. The False Reject error and the False Accept error.

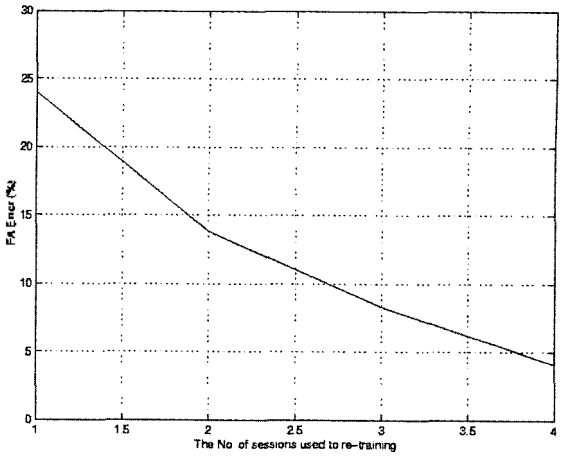

Figure 5. The False Accept error for periodic off-line model re-training with due consideration of the time varying.

ground model normalization. The verification system used the simple left-to-right HMM composed of 3 stats for all phoneme models.

Fig. 4 shows FR (False Reject) error and FA (False Accept) error. Solid line described FA and FR Error of virtual cohort normalization and dash-dotted line is that of background model normalization. As shown in Fig. 4, virtual cohort normalization using f-ratio is lower FA error rate than that of background model normalization respectively. Using the f-ratio is contributed to getting characteristic of speaker's voice, but on the other hand, it have an adverse effect on capturing the variance in speaker's voice over time. Therefore periodic off-line model adaptation or re-training was employed to adapt to the variation of speaker's voice after long time passed. Fig. 5 shows a significant decline of FA Error for periodic off-line model re-training with due consideration of time varying. The mVprotek system used a way of time switch for periodic off-line model re-training.

As a result of this experiment, we got the lower EER (Equal Error Rate) as shown in the following table when mVprotek system employed virtual cohort normalization with F-ratio and periodic off-line speaker model adaptation.

In addition to above result of experiment, the mVprotek improved the performance of speaker verification through the smart confirmation technique that is asking a customer his private question.

Table 1. The EER of mVprotek system.

Methods	EER (%)
VC normalization using f-ratio	0.47
Background Model normalization	4.24

VI. Conclusion

In this paper, we applied speaker verification for authentication on wireless internet m-commerce. The speech signal transferred to commerce server as packetized data through embedded EVRC codec by considering Bandwidth. For improving speaker verification rates, we considered two methods. (1) Scenario: private Q&A (2) Speaker verification (SV) algorithm: we use the virtual cohort model and the f-ratio weighted cepstrum. Normalizing the score with f-ratio exhibit superior performance than conventional background normalization technique. We developed 3 kind of clients, WAP, PDA and PC. It also applied to m-commerce on SK Telecom's CDMA network.

Acknowledgment

Thanks to JO Jung, SM Sohn and HJ Chun. Without their unfailing helping, this work could not here been accomplished.

References

1. Y. J. Kyung, et al., Development of authentication technology on m-Commerce, Final Report, SK Telecom, Seoul, January 2001.
2. Y. J. Kyung, et al., "User verification method and system by using voice for use in electronic commerce," Document of patent pending, 10-2000-0070696, Korea, 2000.
3. S. M. Sohn, and Y. J. Kyung, et al., "A method of distinguishing voice from noise of portable mobile terminal," Document of patent pending, 10-2000-0067531, Korea, 2000.
4. T. Isobe, and J. Takahashi, "A new cohort normalization using local acoustic information for speaker verification," *Proc. of ICASSP*, 1999.
5. S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition," *Electronic Communications* A **57**, 34-42, 1974.
6. Y. J. Kyung, et al., "The mVprotek: m-commerce voice verification system," *Proc. of Eurospeech*, 2001.

[Profile]

• Youn-Jeong Kyung

Youn Jeong Kyung received the B.S. degree and M.S. degree in Computer Science from DongDuk Women's University, Korea, in 1984, 1992 and the Ph.D. in EECS from KAIST, Korea, in 2000. Since 2000 she has been a researcher in Platform R&D center of SK Telecom, Seoul, Korea. Since 2003 she also has been a resarcher in CS, Tokyo Institute of Technology, Tokyo, Japan.