한국ITS학회논문지
제10권, 제1호 (2011년 2월)
pp.84~90

텔레매틱스 환경에서 화자인증을 이용한 VoIP기반 음성 보안통신

VoIP-Based Voice Secure Telecommunication Using Speaker Authentication in Telematics Environments

김 형 국* **신 동****
(Hyoung-Gook Kim) (Dong Shin)

요 약

본 논문은 텔레매틱스 환경에서 문장독립형 화자인증을 이용한 VoIP 음성 보안통신기술을 제안한다. 보안통신을 위해 송신측에서는 화자의 음성정보로부터 생성된 공개키를 통해 음성 패킷을 암호화하여 수신측에 전송함으로써 중간자 공격에 대항한다. 수신측에서는 수신된 암호화된 음성패킷을 복호화한 후에 추출된 음성 특징과 송신측으로부터 수신받은 음성키를 비교하여 화자인증을 수행한다. 제안된 방식에서는 Gaussian Mixture Model(GMM)-supervector를 Bayesian information criterion (BIC) 방식과 Mahalanobis distance (MD) 방식을 이용한 Support Vector Machine (SVM) 커널에 적용하여 문장독립형 화자인증 정확도를 향상시켰다.

Abstract

In this paper, a VoIP-based voice secure telecommunication technology using the text-independent speaker authentication in the telematics environments is proposed. For the secure telecommunication, the sender's voice packets are encrypted by the public-key generated from the speaker's voice information and submitted to the receiver. It is constructed to resist against the man-in-the middle attack. At the receiver side, voice features extracted from the received voice packets are compared with the reference voice-key received from the sender side for the speaker authentication. To improve the accuracy of text-independent speaker authentication, Gaussian Mixture Model(GMM)-supervectors are applied to Support Vector Machine (SVM) kernel using Bayesian information criterion (BIC) and Mahalanobis distance (MD).

Key words : Text-independent speaker verification, bayesian information criterion, Mahalanobis distance, public-Key infrastructure, diffie-hellman algorithm

Ⅰ. 서 론

국내외 유·무선 통신망이 모두 IP망으로 진화하면서 무선 IP망을 통해 텔레매틱스 서비스를 제공하는 시도가 이루어지고 있다. 이를 통해 길안내, 위치정보 또는 교통정보 등을 제공했던 운전자 중

† 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구사업임(2010-0011674).
* 주저자 및 교신저자 : 광운대학교 전파공학과 부교수
** 공저자 : 광운대학교 전파공학과 석사과정
† 논문접수일 : 2010년 11월 23일
† 논문심사일 : 2011년 12월 28일
† 게재확정일 : 2011년 1월 13일

심의 초기 텔레매틱스 서비스는 인터넷을 기반으로 한 정보제공, 음악, 영화감상, 그리고 인터넷 전화 등과 같은 승객 중심의 멀티미디어 서비스로 확장되고 있으며, 기업에서는 이러한 멀티미디어 서비스에 적합한 안드로이드 기반의 IP 텔레메틱스 모바일 단말장치 개발이 진행 중이다. 이로 인해 자동차의 승객은 차내의 IP 텔레매틱스 모바일 단말장치를 이용해 다양한 멀티미디어 서비스를 제공받을 수 있을 뿐만 아니라, 인터넷전화를 통해 자유롭게 통화를 할 수 있게 된다.

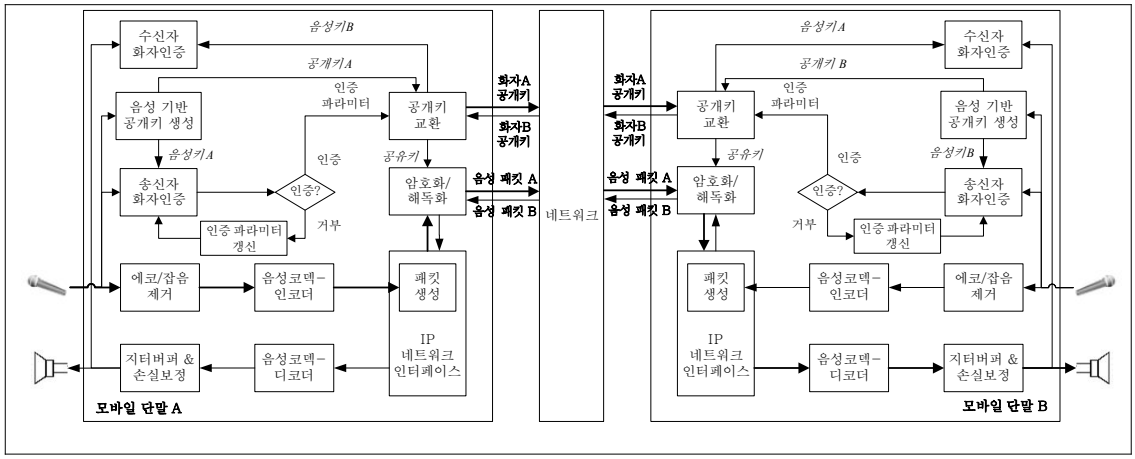
VoIP (Voice over Internet Protocol)는 전 세계 어디에서나 쉽게 접속할 수 있는 IP 기반 네트워크를 통한 음성 통신용 기술로서 현재 빠른 속도로 텔레매틱스 멀티미디어 서비스에 적용되고 있을 뿐만 아니라, 재해·재난 발생 시의 인명구조, 교통관리, 피해복구 등을 경찰들이 효과적으로 수행하기 위한 음성 보안통신에도 적용될 수 있다. 그러나 모바일 VoIP는 기존 전화망을 사용하지 않고 인터넷을 통해 통신을 하기 때문에 통화비가 저렴한 장점이 있는 반면에 보안에 취약한 단점을 갖고 있다. 이러한 모바일 VoIP 통신 시에 발생할 수 있는 패킷 스니핑 및 RTP (Real-time Transport Protocol) 조작을 통한 3자의 도청 및 해킹 공격을 차단하기 위해 RTP 암호화기반의 공개키를 이용한 SRTP (Secure RTP)[1]와 같은 신뢰성 있는 서비스 기술이 개발되었다. 그러나 Diffie-Hellman (DH) 암호 알고리즘[2] 기반의 SRTP 키 교환 기술에서도 세션 가로채기, MITM (Man-in-the-Middle)를 통한 위협이 존재하기 때문에, 지문, 홍채, 망막, 얼굴 특징, 음성 등의 사용자 자신만의 고유한 특성을 반영하는 생체기반 본인인증 방식[3]을 보안통신에 적용하기 위한 연구가 진행되어 오고 있다.

기존의 공개키 생성 방법은 제3의 인증기관을 통해 공개키를 생성하여 전달하고 사용자는 개인키를 통해 공개키로 암호화된 데이터를 해독화한다. 이러한 방식은 제3의 인증기관이 개입해야 되는 번거로운 단점을 갖고 있다.

본 논문에서는 음성특징을 특정비트만큼 변환하여 공개키로 사용하는 방법[4]을 적용하여 제3의 인증기관을 거치지 않고 음성 보안통신을 수행하는 방식을 소개한다.

Ⅱ. 모바일 VoIP 음성 보안통신 시스템

<그림 1>은 본 논문에서 제안한 텔레메틱스 모바일 VoIP 음성 보안통신 시스템의 전체적인 구조를 나타낸다. 각 자동차의 모바일 단말 A, B에서 통화 전에 각 화자의 음성정보로부터 음성키를 생성하고, 생성된 음성키를 기반으로 미리 공개키를 생성한다. 호 연결을 요청할 때 단말 A에서는 단말



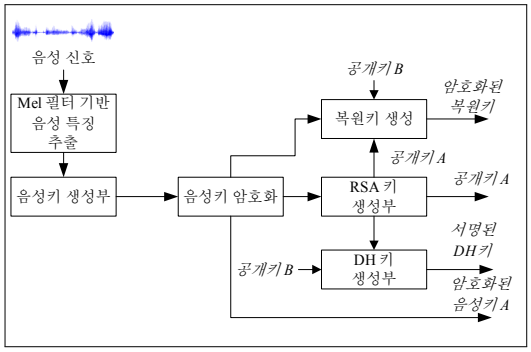
〈그림 1〉 모바일 VoIP 음성 보안통신 시스템 구성도
〈Fig. 1〉 Block diagram of mobile VoIP voice secure telecommunication system

A에서 생성된 음성 공개키(그림의 화자A 공개키)를 단말 B측에 전달한다. 이에 대응하여 단말 B는 호에 대한 응답으로써 단말 B에서 생성된 음성 공개키(그림의 화자B 공개키)를 단말 A에 전달하며, 이 과정에서 음성통신에 사용될 부호화/복호화 방식(G. 711, G. 729, etc)을 설정한다. 각 단말기에 교환된 상대측 공개키는 미리 각 단말에 저장된 공개키와 조합되어 음성패킷을 암호화/복호화 할 수 있는 공유키를 생성한다. 그 이후에 통화를 위해 단말의 외부 입력장치로부터 음성 신호가 입력되면 에코/잡음 제거를 거쳐 호 설정 과정에서 선택된 부호화 방식을 통해 인코더에서 음성신호가 부호화 된다. 이에 대응하여 부호화된 음성신호는 패킷으로 분할되고, 각 음성패킷은 공유키를 통해 암호화되어 상대방 단말에 전달된다. 단말 B로부터 수신된 암호화된 패킷은 공유키를 통해 해독화되고, 해독화된 음성패킷은 설정된 부호화 방식에 대응되는 하나의 복호화 방식을 선택하여 음성신호로 복호화된다. 복호화된 음성신호는 외부 출력장치로 전달되기 이전에, 화자인증을 수행하게 된다. 즉, 복호화된 음성신호로부터 화자 B의 음성키를 추출하고 수신된 공개키로부터 복원된 음성키 B와 비교하는 화자인증을 통해 동일 인물인 경우에 한하여 복호화된 음성신호를 외부 출력장치로 전달하여 송수신간의 통화를 허락한다. 이와 마찬가지로 단말 B에서도 화자A에 대한 화자인증을 수행하여 두 단말에서 모두 화자인증이 성공하였을 때 통화가 지속되게 된다.

1. 음성 기반 공개키 생성

음성 기반 공개키 생성부는 화자의 음성신호로부터 변조 음성키를 생성하고, 변조 음성키의 암호화를 기반으로 이중으로 보안된 공개키와 암호화된 복원키를 생성한다. 그리고 서명된 DH키는 공개키로부터 생성된다.

(Step1) Mel 필터기반 음성특징 추출: 단말에 입력된 8kHz sampling rate를 갖는 연속된 음성신호로부터 분할된 20ms 단위의 음성신호 프레임의 에너지 레벨 점검을 통해 선별된 음성구간을 고속 푸리



〈그림 2〉 음성키 기반 공개키 생성부 구성도
〈Fig. 2〉 Block diagram of voice based public key generator

에 변환과 23개의 멜 스케일링필터로 통과시키고, 로그변환과 Discrete Cosine Transform을 거쳐 프레임당 13차의 MFCC 계수와 13차의 delta-cepstral 계수를 포함하여 총 26차로 구성된 음성특징을 추출한다. 획득된 화자의 음성특징을 기반으로 MAP adaptation을 수행하여 화자의 유사도를 최대화하는 모델 파라미터인 화자 GMM supervector를 형성한다.

(Step2) 변조 음성키 생성 및 음성키 암호화: 3자의 해킹에 대비하기 위해 입력음성으로부터 생성된 GMM supervector X(d,l)으로 부터 변조할 Z개의 각 프레임 별 차원의 특정값 μlm을 계산한다. 계산된 특정값 μlm은 식(1)에 적용되어 원래의 음성특징값을 변조시킨다.

$$\hat{X}(d,l) = \begin{cases} X(d,l) + \mu_l & \text{if } l = lm \\ X(d,l) & otherwise \end{cases} \quad (1)$$

여기서 X는 GMM supervector, d는 특징차수, l은 프레임 지수, lm은 프레임 지수 l 내에서의 임의의 프레임 지수이다.

암호화된 오디오 특징은 각 차수별 크기를 3자리의 숫자로 감소시킨 후에, 부호는 1비트로 표현하고, 3자리의 숫자는 10비트 2진수로 표현하여 하나의 십진수를 11비트의 이진수로 나타낸다. 위 과정을 통해 변조된 음성키가 다음과 같이 생성된다.

$$VK_{AM} \approx VK_{AM}(d,l) = (\hat{X}(d,l))_{(2)} \quad (2)$$

(Step3) RSA키 생성: 암호화된 음성키에 DH기반의 RSA알고리즘[2]을 적용하여 음성 공개키를 생성한다. 먼저, 특정값 μ_{lm} 을 기반으로 식(5)와 같이 하나의 연속되는 수 *Z*를 생성하고, *Z*보다 작은 소수 *e*를 생성하여 RSA 공개키의 하나로 사용한다.

$$Z = \lfloor \mu_{lm1}\mu_{lm2}\mu_{lm3}\mu_{lm4}\mu_{lm5}\mu_{lm6} \rfloor \quad (3)$$

$$e = Z - \varepsilon \quad (4)$$

식에서 ε 는 *Z*와 소수 *e* 사이의 차이, *e*는 RSA 공개키의 하나로 사용되는 소수이다.

그 후에, 두 개의 큰 임의의 다른 소수 *r*와 *q*를 생성하여 *n=rq*를 계산함으로써, RSA 공개키와 개인키는 각각 식(5), 식(6)을 통해 생성된다.

$$PK_A^+ = [e,n] \quad (5)$$

$$PK_A^- = [d,n] \quad (6)$$

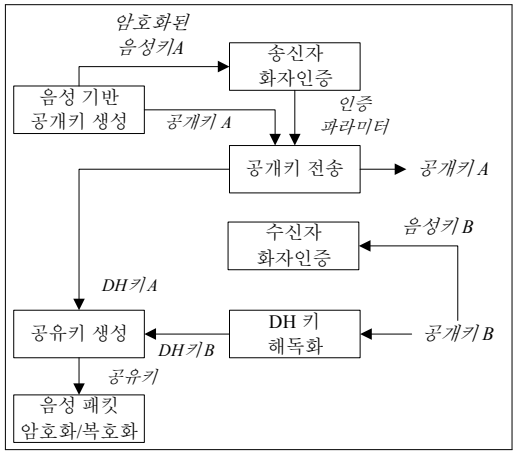
식(6)의 *d*는 1<*d*<$\phi(n)$의 조건을 만족해야 하며, $\phi(n)$은 $\phi(n)=(p-1)(q-1)$를 통해 계산된다. 그리고 공개키 *e*와 $\phi(n)$사이에는 최대공약수 gcd(*e*,ϕ)=1이어야 하며, *e*와 $\phi(n)$를 특별한 정수 RSA 개인키 *d*에 적용하기 위해서는 식(7)가 성립되어야 한다.

$$(e \times d) \bmod \varphi(n) = 1 \quad (7)$$

(Step4) DH키 생성: 생성된 RSA 키 쌍을 식(10)에 적용하여 DH키를 생성한다.

$$DH\,PK_A(g,x,p) = g^x \bmod p \quad (8)$$

식(8)에서 g와 *p*는 각 화자에게 알려져 있는 DH키를 구성하는 정수이고, RSA 개인키 *d* 를 적용하여 화자 A의 DH 개인 키는 $x=d_A$, 화자 B의 DH 개인 키는 $y=d_B$ 로 나타낼 수 있다. *g*, *p*, *x*로부터 단말 A의 DH 공개키를 생성한다.



〈그림 3〉 공개키 교환
〈Fig. 3〉 Block diagram of public key exchange

생성된 DH 공개키와 수신된 RSA 공개키 PK^+_B를 식(9)에 적용하여 서명된 DH 키를 구한다.

$$Sign(DH\,PK_A)(g,x,p,PK_B^+) = Sign_{PK_B^+}(g^x \bmod p) \quad (9)$$

(Step5) 복원키 생성: 암호화된 음성키를 복원하기 위해 ε 와 RSA 공개키를 통해 복원키를 식(10)와 같이 생성한다.

$$SK_A = [\epsilon, PK_A^+] \quad (10)$$

생성된 복원키는 수신된 단말 B의 공개키를 통해 암호화된다.

$$SK_{AE} = (SK_A)_{PK_B^+} \quad (11)$$

위 과정을 통해 음성기반 공개키인 음성키, RSA 공개키, DH키, 복원키가 생성된다.

2. 공개키 교환

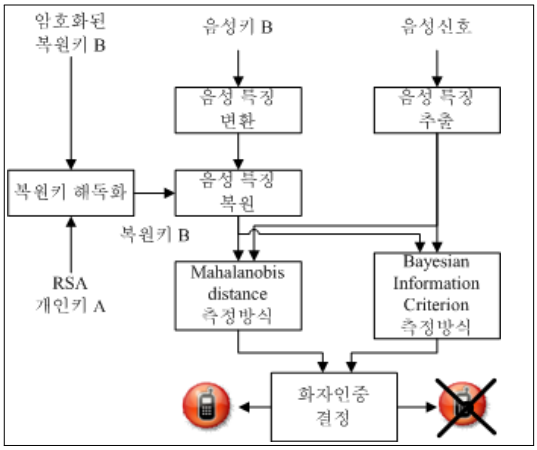
음성 기반 공개키 생성과정을 통해 생성된 공개키는 <그림 3>의 과정을 통해 단말 A와 B간에 서로 교환된다. 공개키 교환부에서는 통화 전에 송신

자가 통화 환경변화에 따른 화자인증 파라미터를 갱신하기 위해 송신자 화자인증을 수행하여 파라미터가 갱신되고, 다이얼링이 시작되면 공개키를 서로 교환하게 된다. 이 때 전송되는 공개키는 RSA 공개키를 먼저 전송하고, 수신된 RSA 공개키를 통해 암호화된 음성키, 서명된 DH키를 전달하게 된다. 수신된 공개키 중 DH키인 $g^y \bmod p$는 자신의 RSA 개인키로 해독화되고 자신의 DH키 $g^x \bmod p$와 결합하여 음성패킷을 암호화하고 해독화할 수 있는 공유키인 $(g^y)^x \bmod p$를 생성한다.

3. 화자인증

<그림 4>는 단말 A에 구현된 화자인증 구성도로서 다음과 같이 수행된다.

수신된 단말 B의 암호화된 복원키는 단말 A의 RSA 개인키를 통해 복원키 B로 해독화되고 음성특징 복원부로 전달된다. 수신된 변조된 음성키 VK_{BM}은 음성특징으로 변환된 후에, 복원키 B에 의해 화자인증에 필요한 음성 특징값으로 복원된다. 복원키에서 암호화된 특정값 μ_{lm}으로 구성된 Z를 복원하기 위해 RSA 공개키를 구성하는 e와 복원키 c를 사용하고, 변조된 특정값 μ_{lm}으로 구성된 Z는 다음과 같이 암호화된 음성특징과 결합되어 음성특징값을 해독한다.



〈그림 4〉 화자 인증부
〈Fig. 4〉 Block diagram of speaker verification

$$X(d,l)=\begin{cases}\hat{X}(d,l)-\mu_l \text{ if } l=lm\\ \hat{X}(d,l) \qquad otherwise\end{cases} \tag{12}$$

문장독립형 화자인증의 정확도를 높이기 위해서 본 논문에서는 GMM supervector를 SVM 커널에 적용하여 화자를 인식하는 방법을 적용한다. SVM은 최적의 하이퍼분리면을 찾아서 화자와 비화자의 두 범주를 분류하는 기법으로서 식(13)와 같이 표현된다.

$$f(X)=\sum_{j=1}^{J}\alpha_j t_j K(X_{a_j},X_{b_j})+d \tag{13}$$

식(13)에서 α는 Lagrangian multiplier, d는 바이어스 $K(,)$는 커널함수로서, 본 논문에서는 Bayesian Information Criterion(BIC)[5] 방식과 Mahalanobis Distnace (MD)[6]를 SVM 커널에 적용하여 각각 획득된 화자인증 결과를 결합하여 화자인증을 결정한다.

입력되는 화자 a와 화자 b의 GMM supervectors의 유사성을 결정하는 SVM-BIC 커널은 식(14)와 같이 정의되며, SVM-MD 커널은 식(15)와 같이 정의된다.

$$\begin{aligned}&\Psi_{BIC}(X_a\|X_b)\\&=R(X_a\|X_b)-\lambda\frac{1}{2}\left(p+\frac{1}{2}p(p+1)\right)\log N_X,\end{aligned} \tag{14}$$

$$\begin{aligned}&R(X_a\|X_b)\\&=\frac{N_X}{2}\log|\textstyle\sum X|-\frac{N_{X_a}}{2}\log|\textstyle\sum X_a|-\frac{N_{X_b}}{2}\log|\textstyle\sum X_b|\end{aligned}$$

여기서 p는 특징 벡터의 차수, λ는 문턱값을 나타내며, X_a, X_b는 각각 수신된 음성 지문으로부터 복원된 음성 특징, 수신된 음성패킷으로부터 추출된 음성 특징, Σ_X, Σ_{Xa}, Σ_{Xb} 는 각각 전체 구간에 대한 공분산, X_a, X_b의 공분산 행렬이다. 그리고 N 은 음성 특징의 개수이다.

$$\begin{aligned}&\Psi_{Mahal}(X_a\|X_b)\\&=\sqrt{\sum_{i=1}^{p}(X_a(i)-X_b(i))^T(\textstyle\sum_i^{(a)}+\sum_i^{(b)})^{-1}(X_a(i)-X_b(i))}\end{aligned} \tag{15}$$

SVM-BIC 커널과 SVM-MD 커널은 식(16)에 의해 결합된다.

$$\Gamma_{Fusion}(X_a \| X_b) = \begin{cases} 1 & , \text{if } \Gamma_{BIC} = 1 \text{ and } \Psi_{MD} \le \delta_{Maha1} \\ 1 & , \text{if } \Gamma_{BIC} = 0 \text{ and } \Psi_{MD} \le \delta_{Maha2} \\ 0 & , otherwise \end{cases} \quad (16)$$

여기서

$$\Gamma_{BIC}(X_a \| X_b) = \begin{cases} 1 & , \text{if } \Psi_{BIC} < 0 \\ 0 & , otherwise \end{cases} \quad (17)$$

$\Gamma_{BIC}(X_a \| X_b)$는 SVM-BIC 커널로부터 계산된 BIC 커널 점수이며, BIC 커널 점수는 0을 기준으로 0보다 작은 경우 동일인물로 판단하고, 0보다 큰 경우 다른 인물로 판단한다.

$$\Gamma_{Maha}(X_a \| X_b) = \begin{cases} 1 & , \text{if } \Psi_{MD} \le \delta_{Maha1} \\ 0 & , otherwise \end{cases} \quad (18)$$

$\Gamma_{Maha}(X_a \| X_b)$는 SVM-MD 커널로부터 계산된 MD 커널 점수이며, MD 커널 점수는 문턱값 δ_{Maha1}을 기준으로 작으면 동일인물로 크면 다른 인물로 판단한다.

Ⅲ. 실험 및 결과 고찰

본 논문에서는 제안하는 화자인증 알고리즘을 Windows Mobile 6.1기반의 SCH-M480 모바일 기기(Marvel PXA310 625MHz 프로세서, 128MB RAM메모리, 256MB ROM메모리)에 구현하고, 구현된 모바일 단말기를 자동차 내에 장착하여 두 단말간의 음성보안통신에 대한 성능을 측정하였다.

실험에 사용된 음성데이터는 거리, 고속도로 등에서 자동차 주행 시에 녹음된 배경잡음과 화자음성파일(남자 30명, 여자 30명으로 구성된 총 60명의 화자가 발성한 텍스트 내용이 다른 2분길이의 각 10개의 문장 8kHz/8 비트의 PCM 방식으로 녹음)을 혼합하여 잡음환경에 노출된 음성 DB (SNR 10~20 dB)로 구성되었으며, 음성신호에 대해 20 ms의 프레임 크기의 해밍윈도를 이용하여 13차의 MFCC 계수와 13차의 delta-cepstral 계수를 포함하여 총 26차로 구성된 음성특징을 추출하여 공개키 생성과 화자인증에 적용하였다.

화자 및 비화자 모델링은 음성 DB의 각 화자로부터 추출된 음성 특징값을 2048 개의 가우시안 믹쳐를 갖는 GMM-UBM (Universal Background Model) 방식에 적용하여 2325 GMM supervector를 획득하였다. 그리고 추출된 GMM supervector와 SVM 배경을 사용하여 학습을 수행하였다.

화자인증 테스트는 SNR 10~20dB 사이의 자동차 잡음환경에서 총 60명의 화자(남자 30명, 여자 30명)가 학습에 사용되지 않은 6초 길이의 독립된 20개의 문장을 발성하여 실시간으로 VoIP 음성통신을 수행함으로써 성능을 측정하였다.

본 논문에서 제안된 방식을 기반으로 실시간 보안통신적용 화자인증 실험을 실시한 결과, 화자인증률 93.2%로서 6.8%의 화자인증 오류를 나타내었는데, 이 화자인증 결과는 일반적인 GMM 방식의 화자인증률(82%) 보다 11.2%, GMM supervector를 SVM Kullback-Leibler distance 커널방식[7]에 적용한 화자인증률(87.8%)보다 5.4%, GMM supervector를 SVM-BIC 커널방식에 적용한 화자인증률(88.3%)보다 4.9% 적은 화자인증 오류결과를 나타냄을 알 수 있었다. 발생한 화자인증오류는 변동되는 VoIP 네트워크 상황에 따라 수신된 음성에 패킷 손실에 의해 발생하였으며, 또한 통화시의 갑작스런 음성의 템포와 강도 등의 변화로 인해 화자인증 정확도가 낮아짐을 알 수 있었다.

Ⅳ. 결 론

본 논문에서는 무선 모바일 VoIP 환경에서 보안통신을 위한 문장독립형 화자인증 시스템을 제안하였다. 제안한 시스템은 모바일기기에서 93.2%의 높은 화자 인증률을 보여주었다.

향후 계획으로는 다양한 잡음환경, 다양한 채널과 대용량 음성 DB기반에서 제안된 화자인증 방식에 Joint Factor Analysis를 결합함으로써 화자인증 정확률을 향상시키기 위한 연구를 수행할 예정이다.

참 고 문 헌

[1] B. H. Song, K. S. Chung and Y. T. Shin, "SRTP: TCP-friendly congestion control for multimedia streaming," *Lecture Notes in Computer Science Springer-Verlag Press,* vol.2344, pp.529~538, Sep. 2002.

[2] M. E. Hellman, "An overview of public key cryptography." *IEEE Communications Magazine*, pp.42~49, May 2002.

[3] D. J. Kim and K. S. Hong, "Multimodal biometric authentication using teeth image and voice in mobile environment," *IEEE transactions on Consumer Electronics*, vol.54, no.4, pp.1790~1797, Nov. 2008.

[4] M. R. Enayah and A. Samsudin, "Securing telecommunication based on speaker voice as the public key," *IJCSNS*, pp.201~210, 2007.

[5] P. Delacourt and C. J. Wellekens, "DISTBIC: a speaker-based segmentation for audio data indexing," *Elsevier Speech Communication*, vol.32, pp.111~126, Sep. 2000.

[6] R. D. Maesschalck, D. Jouan-Rimbaud and D. L. Massart, "Tutorial the Mahalanobis distance," *Elsevier Chemometrics and Intelligent Laboratory systems*, vol.50, pp.1~18, 2000.

[7] W. M. Campbell, D. E. Sturim and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol.13, no.5 pp.308~311, May 2006.

저자소개

김 형 국 (Kim, Hyoung-Gook)

2007년 3월 ~ 현 재 : 광운대학교 전파공학과 부교수
2005년 4월 ~ 2007년 2월 : 삼성종합기술원 수석연구원
2002년 8월 ~ 2005년 3월 : 독일 베를린 공과대학교 Assistant Professor
1999년 1월 ~ 2002년 7월 : 독일 SIEMENS/Cortologic AG 책임연구원

신 동 (Shin, Dong)

2009년 3월 ~ 현 재 : 광운대학교 전파공학과 석사과정
2009년 : 광운대학교 전파공학과 공학사