논문 2008-45CI-5-21

이동환경에서 치열영상과 음성을 이용한 멀티모달 화자인증 시스템 구현

(An Implementation of Multimodal Speaker Verification System using Teeth Image and Voice on Mobile Environment)

김 동 주*, 하 길 람*, 홍 광 석**

(Dong-Ju Kim, Kil-Ram Ha, and Kwang-Seok Hong)

요 약

본 논문에서는 이동환경에서 개인의 신원을 인증하는 수단으로 치열영상과 음성을 생체정보로 이용한 멀티모달 화자인증 방법에 대하여 제안한다. 제안한 방법은 이동환경의 단말장치중의 하나인 스마트폰의 영상 및 음성 입력장치를 이용하여 생체정보를 획득하고, 이를 이용하여 사용자 인증을 수행한다. 더불어, 제안한 방법은 전체적인 사용자 인증 성능의 향상을 위하여 두 개의 단일 생체인식 결과를 결합하는 멀티모달 방식으로 구성하였고, 결합 방법으로는 시스템의 제한된 리소스를 고려하여 비교적 간단하면서도 우수한 성능을 보이는 가중치 합의 방법을 사용하였다. 제안한 멀티모달 화자인증 시스템의 성능평가는 스마트폰에서 획득한 40명의 사용자에 대한 데이터베이스를 이용하였고, 실험 결과, 치열영상과 음성을 이용한 단일 생체인증 결과는 각각 8.59%와 11.73%의 EER를 보였으며, 멀티모달 화자인증 결과는 4.05%의 EER를 나타냈다. 이로부터 본 논문에서는 인증 성능을 향상하기 위하여 두 개의 단일 생체인증 결과를 간단한 가중치 합으로 결합한 결과, 높은 인증 성능의 향상을 도모할 수 있었다.

Abstract

In this paper, we propose a multimodal speaker verification method using teeth image and voice as biometric trait for personal verification in mobile terminal equipment. The proposed method obtains the biometric traits using image and sound input devices of smart-phone that is one of mobile terminal equipments, and performs verification with biometric traits. In addition, the proposed method consists the multimodal-fashion of combining two biometric authentication scores for totally performance enhancement, the fusion method is accompanied a weighted-summation method which has comparative simple structure and superior performance for considering limited resources of system. The performance evaluation of proposed multimodal speaker authentication system conducts using a database acquired in smart-phone for 40 subjects. The experimental result shows 8.59% of EER in case of teeth verification, 11.73% in case of voice verification, and the multimodal speaker authentication result presented the 4.05% of EER. In the experimental result, we obtain the enhanced performance more than each using teeth and voice by using the simple weight-summation method in the multimodal speaker verification system.

Keywords: multimodal biometrics, teeth verification, voice verification

I. 서 론

정보사회의 발전과 더불어 인간과 컴퓨터사이의 인터페이스 기술이 점점 부각되면서 생체인식 기술에 대한 연구가 활발히 진행되고 있다. 또한 개인의 정보의 중요성이 대두되면서 정보의 보안 및 관리, 개인의 신

* 학생회원, ** 정회원, 성균관대학교 정보통신공학부
(School of Information and Communication Engineering, Sungkyunkwan University)

※ 이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-0889-000).

접수일자: 2007년11월12일, 수정완료일: 2008년8월29일

분 증명을 위한 방법으로 인간의 얼굴, 음성, 홍채, 지문, 서명, 정맥과 같은 생체 정보를 독립적으로 이용하는 기술이 수행되어 왔다. 하지만 이러한 단일 생체인식 방법은 각 생체 정보마다 가지는 취약점으로 인하여, 인식 성능의 저하를 가져온다. 이와 같은 단일 생체인식 기술의 한계를 보완하고 인증 성능의 향상과 신뢰도를 높이기 위하여 최근에는 얼굴과 음성, 지문과 얼굴, 홍채와 지문, 정맥과 홍채 등과 같이 두 가지 이상의 생체정보를 이용하는 멀티모달 방식이 활발히 연구되고 있다[1]. 이와 더불어 최근에 널리 보급되어 활용되고 있는 이동단말 장치는 센서, 프로세서, 커뮤니케이션, 인터페이스, 보안 등의 핵심 요소 기술을 요구하고 있으며, 특히 이동 네트워크 상황에서 정보가 도처에 존재함에 따른 정보보안 및 프라이버시의 중요성이 강조되고 있다. 이러한 이동단말 환경에서 정보보안의 취약성을 극복하기 위해서는 기밀성 (Confidentiality), 인증성 (Authentication), 무결성 (Integrity) 등이 요구되며, 인증수단으로 생체정보를 이용하는 방안이 활발하게 적용되고 있다[2~3].

이에 본 논문에서는 이동단말 장치에서 개인의 신원을 인증하는 수단으로 치열영상과 음성을 생체정보로 이용한 멀티모달 화자인증 방법에 대하여 제안한다. 치열영상과 음성은 영상 및 음성 입력장치를 이용하여 획득되는 생체정보로서, 최근의 이동단말 환경에서 이러한 장치들은 기본사양으로 자리매김 되어가고 있다. 그러므로 제안한 방법은 생체정보 입력을 위한 추가의 장치가 부가되지 않아 저가의 비용으로 시스템을 구축 할 수 있는 장점을 갖는다. 제안한 멀티모달 화자인증 시스템은 치열인증과 음성인증 모듈로 구성된다. 치열영상을 이용한 생체인식 방법은 처음으로 LDA (Linear Discriminant Analysis)와 NN (Nearest Neighbor) 분류기를 사용한 방법이 제안되었고[4], 더불어 PCA (Principal Component Analysis)와 NN 분류기를 이용한 성능 개선에 관한 연구[5~6]와 다양한 영상인식 알고리즘에 대한 성능비교에 관한 연구를 찾아볼 수 있다[7]. 치열영상을 이용한 생체인식 방법의 성능평가 연구[7]에 따르면, 최적의 치열인식 성능을 보이는 방법은 2D-DCT (Two Dimensional Discrete Cosine Transform)와 EHMM (Embedded Hidden Markov Model) 알고리즘을 사용한 경우이다. 이에 본 논문에서는 치열인증 모듈에 2D-DCT와 EHMM을 이용한 방법을 적용하였다. 이와 더불어 치열 영상을 획득하는 과정에서 사용자에게 /이/ 음을 발성하도록 부과함으로써, 획득한 음성을 음성인증에 이용하였다. 음성인증에는 특징벡터로 MFCC (Mel Frequency Cepstral Coefficient)와 피치를 사용하였고, GMM (Gaussian Mixture Model) 알고리즘으로 사용자의 음성을 모델링하였다.

대부분의 멀티모달 생체인식 시스템은 각 생체정보들의 결합 방식에 따라 특징 단계, 유사도 단계, 결정 단계로 구분될 수 있다. 본 논문에서는 전체적인 인증 성능의 향상을 위하여 치열과 음성의 인증 결과를 유사도 단계에서 결합하였으며, 시스템의 제한된 리소스를 고려하여 비교적 간단하면서도 우수한 성능을 보이는 가중치의 합으로 결합하여 시스템을 구성하였다. 제안한 방법의 성능평가는 스마트폰을 이용하여 구축한 데이터베이스를 이용하여 수행하였다. 데이터베이스는 실제 환경에서 사용자가 움직이지 않는 상태를 전제로 하여 구축되었으며, 전체적으로 40명에 대한 800개의 치열영상과 음성으로 구성된다.

Ⅱ. 치열 인증

본 절에서는 치열인증 모듈의 구성에 대하여 기술한다. 치열인증은 크게 입력영상에서 치열영역을 검출하는 모듈, 특징 파라미터 추출 모듈, 그리고 유사도 계산하는 모듈로 구분된다. 입력영상에서의 치열영역을 검출하는 모듈에는 빠른 검출 속도를 보이는 Haar-like feature 기반의 AdaBoost 알고리즘을 이용하며, 치열영상의 특징 파라미터로는 2D-DCT, 치열영상을 위한 모델 학습과 인증에는 EHMM 알고리즘을 사용한다.

1. 치열영역 검출

치열인증을 위한 첫 단계는 이동단말 장치중의 하나인 스마트폰에서 영상을 획득하고, 획득한 영상에서 치열영역을 검출하는 과정이다. 본 논문에서는 치열영역을 검출하기 위하여 최근에 가장 널리 사용되는 Haar-like 특징 기반의 AdaBoost 학습 알고리즘을 사용하였다[8]. AdaBoost 학습 알고리즘은 여러 개의 약한 분류기인 h_j의 선형적인 결합을 통하여 최종적으로 높은 검출 성능을 가지는 강한 분류기인 $H_x(h_1, h_2, h_3 \ldots)$를 그룹화 하는 분류기법이다. Haar-like 특징은 하나의 약한 분류기가 되고, 식 (1)과 같이 정의할 수 있다.

$$h_j(x) = \begin{cases} 1 & \text{if } f_j(x) < \theta_j \\ 0 & otherwise \end{cases} \quad (1)$$

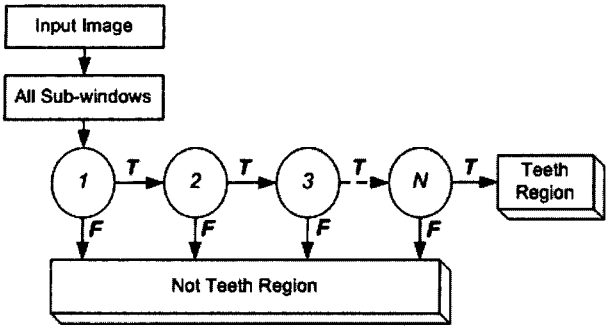


그림 1. 치열 검출을 위한 cascade 구조
Fig. 1. Structure of cascade for teeth region detection.

식 (1)에서 $h_j(x)$는 약한 분류기, f_j는 Haar-like 특징값, θ_j는 문턱치, x는 서브 윈도우의 크기를 의미한다. 그림 1은 탐색 시간을 줄이고 검출 성능을 높이기 위한 cascade 구조로서, 각 단계는 AdaBoost 학습 알고리즘을 통해 추출한 특징값들을 그룹화하여 생성된다.

첫 단계에서는 적은 수의 특징값들을 가지고 일정한 수준의 치열영역 판별 능력을 가진 특징값들로 그룹화하고, 다음 단계에서는 첫 단계에서보다 더 많은 수의 특징값들을 가지고 전 단계보다 세밀한 판별력을 가지는 그룹을 만들어 검출 속도를 향상하게 된다.

2. 특징 파라미터

치열인증에 사용하는 특징 파라미터는 EHMM의 입력으로 사용되는 관측 벡터이다. 다차원의 관측 벡터는 모델의 훈련 및 치열인증 과정에서 많은 계산양의 원인이 된다. 따라서 본 논문에서는 치열영상의 특징 성분을 잘 표현하며 데이터의 중복성을 효율적으로 제거할 수 있는 2D-DCT를 치열영상의 특징 파라미터로 사용하였다. $P\times L$ 크기의 영상에 대한 2D-DCT는 식 (2)과 같이 표현된다.

$$F(u,v)=\frac{2}{\sqrt{PL}}C(u)C(v)\Big[\sum_{i=0}^{P-1}\sum_{j=0}^{L-1}f(i,j)\times \cos\frac{(2i+1)u\pi}{2P}\times\cos\frac{(2j+1)v\pi}{2L}\Big] \quad (2)$$

$$\text{단, } C(u),\ C(v)=\begin{cases}\frac{1}{\sqrt{2}} & for\quad u,v=0\\ 1 & for\quad u,v\neq 0\end{cases}$$

$f(i,j)$: $P\times L$ 영상
$F(u,v)$: 2D-DCT 계수

그림 2는 치열영상의 특징벡터 추출을 위한 윈도우 블록과 중첩의 크기를 나타내고 있다. 폭 W와 높이 H인

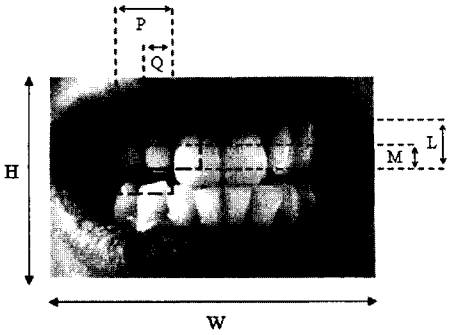


그림 2. 치열영상의 블록 추출
Fig. 2. Block extraction of teeth image.

치열 영상은 특징벡터 추출을 위하여 $P\times L$ 크기의 블록으로 나누어진다. P와 L은 영상의 폭과 높이에 대한 블록 크기로서, 2D-DCT를 계산하는 윈도우 크기이다.

영상의 폭과 높이에 대한 중첩의 크기를 각각 Q와 M이라고 하면, 치열 영상으로부터 추출되는 블록의 개수 T는 식 (3)과 같이 계산된다.

$$T=\left(\frac{W-Q}{P-Q}\right)\times\left(\frac{H-M}{L-M}\right) \quad (3)$$

치열인증에 사용하는 관측벡터 열의 개수는 식 (3)으로 계산되는 블록의 개수와 같으며, 다음과 같은 방법으로 계산된다. $P\times L$ 윈도우는 그림 2의 치열 영상에서 왼쪽에서 오른쪽으로, 위에서 아랫방향으로 스캔한다. 여기에서 이웃한 윈도우 사이의 중첩은 수평방향으로 Q, 수직방향으로 M 만큼 이동하고, $P\times L$ 윈도우를 2D-DCT 변환하여 관측벡터 열을 생성한다.

3. EHMM

본 논문에서는 치열인증을 위한 알고리즘으로 EHMM을 사용하였다. EHMM은 일차원의 HMM을 2차원 구조로 나타내기 위하여 일반화한 방법으로, super-states와 embedded-states의 집합으로 구성된다. 여기에서 각각의 super-state는 하나의 일차원 HMM을 포함하며, 관측확률이 없고 상태천이 확률만을 갖는다. 다음은 EHMM의 구성요소를 나타낸다[9].

Super-state 모델 구성
N_0 : super-state의 개수
Π_0 : super-state 초기상태 분포
A_0 : super-state 상태 천이 확률 매트릭스
Embedded-state 모델 구성
N_1^* : k번째 super-state의 embedded-state 개수

Π_1^k : k번째 super-state의 embedded-state에 대한 초기상태 분포

A_1^k : k번째 super-state의 embedded-state에 대한 상태 천이 확률 매트릭스

B_1^k : k번째 super-state의 embedded-state에 대한 관측 확률 매트릭스

연속 EHMM에서 embedded-state의 관측확률은 식 (4)와 같이 표현된다.

$$b_i^{(k)}(O_{t_0,t_1}) = \sum_{m=1}^{M} c_{im}^{(k)} N(O_{t_0,t_1}, \mu_{im}^{(k)}, U_{im}^{(k)}),\ 1 \le i \le N_1^{(k)} \quad (4)$$

식 (4)에서 O_{t_0,t_1}는 치열영상의 가로축 관측지점 t_0와 세로축 관측지점 t_1에서의 관측 벡터이다. M은 mixture 개수이며, $c_{im}^{(k)}$는 k번째 super-state의 i번째 embedded-state에서 m번째 mixture의 계수이다. $N(O_{t_0,t_1}, \mu_{im}^{(k)}, U_{im}^{(k)})$는 평균벡터 $\mu_{im}^{(k)}$와 공분산 매트릭스 $U_{im}^{(k)}$을 갖는 가우시안 확률밀도 함수를 나타낸다.

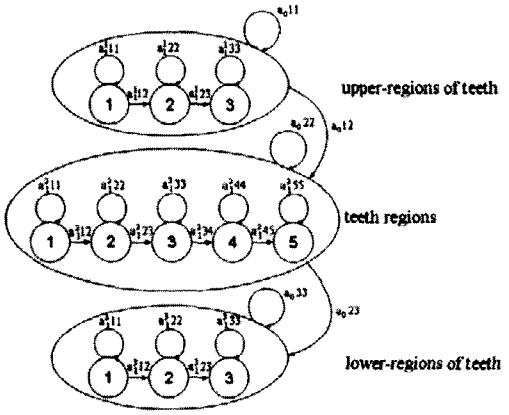


그림 3. EHMM의 모델 구조
Fig. 3. Model structure of EHMM.

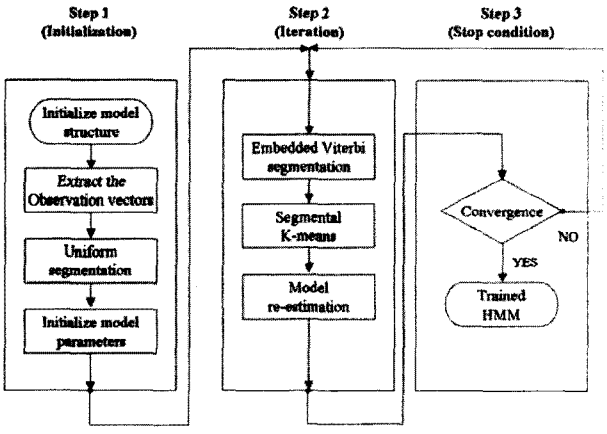


그림 4. 치열모델 학습 블록도
Fig. 4. Block diagram of training teeth model.

그림 3은 super-state의 개수가 3개이고 각각의 super-state가 3개, 5개, 3개의 embedded-state를 포함하는 EHMM 구조를 보여준다. 그림에서 각각의 super-state는 치열의 윗부분, 치열부분, 치열의 아랫부분을 모델링하며, 일차원 HMM을 포함하는 구조를 갖는다.

그림 4는 치열영상을 EHMM 알고리즘을 이용하여 모델을 학습하는 블록도를 나타내고 있다.

치열모델 학습의 처음 과정은 EHMM의 모델 파라미터를 초기화하는 단계이다. 초기화 과정은 super-state와 embedded-state의 상태 개수를 결정하고, 블록추출 및 2D-DCT 관측벡터를 추출하는 단계를 포함한다. EHMM 모델의 파라미터는 치열영상으로부터 추출한 관측벡터를 세로축과 가로축 방향으로 균일하게 분할하여 초기화된다. 초기화된 모델 파라미터들은 doubly embedded Viterbi segmentation 알고리즘과 segmental K-means 알고리즘을 이용하여 재 추정되고, 모델 파라미터들이 수렴할 때까지 재 추정 과정을 반복한다. 모델 학습과정에서 생성한 치열모델은 치열인증 단계에서 사용자 확인에 필요한 임계값 계산에 이용된다.

치열인증은 정해진 사용자를 인증 또는 거절하는 과정으로 사용자의 치열영상과 사전에 학습된 EHMM 모델을 이용하여 수행된다. 그림 5는 치열인증의 블록도로 입력된 치열영상에 대하여 블록 및 관측벡터를 추출하고, doubly embedded Viterbi 알고리즘을 이용하여 확률값을 계산한다. 입력 치열영상에 대하여 얻어진 확률값은 학습과정에서 사전에 계산한 임계값과 비교하고, 이 임계값보다 크면 사용자를 인증하게 되며, 임계값보다 작으면 사용자를 거절하는 구조를 갖는다.

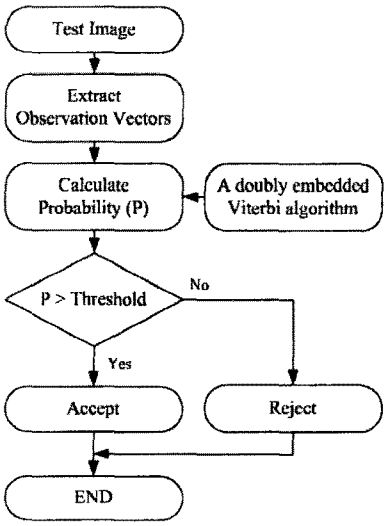


그림 5. 치열인증 블록도
Fig. 5. Block diagram of teeth verification.

Ⅲ. 음성인증

본 논문에서는 치열인증과 더불어 사용자의 음성을 이용한 음성인증을 적용하여 멀티모달 화자인증 시스템을 구성하였다. 음성인증은 치열영상을 획득하는 과정에서 사용자에게 /이/음을 발성하도록 하여, 이 과정에서 획득한 음성신호를 이용한다. 음성인증에는 특징 파라미터로 MFCC와 피치 정보를 결합한 특징 벡터를 사용하고, 음성인증 알고리즘으로 GMM을 사용한다.

1. 특징 파리미터

가. MFCC

MFCC는 음성신호의 대표적인 특징 파라미터로 사람의 귀가 주파수 변화에 반응하게 되는 양상이 선형적이지 않고 로그 스케일과 비슷한 멜 스케일을 따르는 청각적 특성을 반영한 계수 추출 방법이다. 멜 스케일에 따르면 낮은 주파수에서는 작은 변화에도 민감하게 반응하지만 높은 주파수로 갈수록 민감도가 작아지므로 특징 추출시에 주파수 분석 빈도를 이에 맞는 특성에 맞추는 방식이다. MFCC 특징 파라미터를 추출하는 과정은 순차적으로 다음과 단계로 수행되며, 본 논문에서는 MFCC 13차를 음성인증을 위하여 사용하였다.

성문신호의 영향을 최소화하여 음성과 잡음의 구분을 뚜렷하게 하기 위해 고주파 성분을 강조해 주는 Pre-emphasis 과정이 처음으로 수행된다.

음성 신호를 구간으로 나누는 과정에서 주파수 왜곡 현상을 최소화하기 위하여, 윈도우를 신호 성분에 곱하여 분석한다. 본 논문에서는 일반적으로 널리 사용되는 Hamming 윈도우를 사용한다.

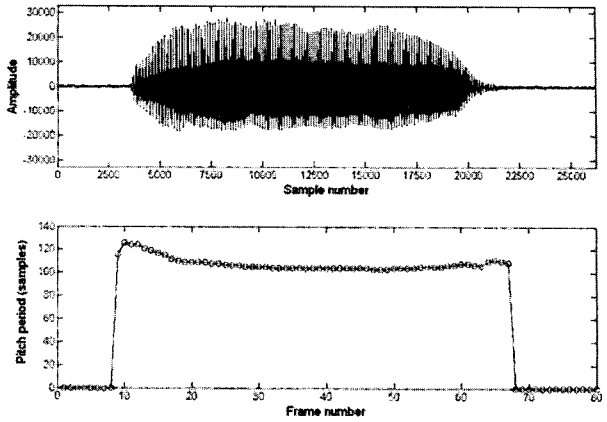
분석구간의 음성신호에 주파수 변환을 취하여 음성신호의 스펙트럼을 구한다.

멜 스케일에 맞춘 삼각 필터뱅크를 대응시켜, 각 밴드에서의 스펙트럼 크기의 합을 계산하고, 필터뱅크의 출력에 로그를 적용한다.

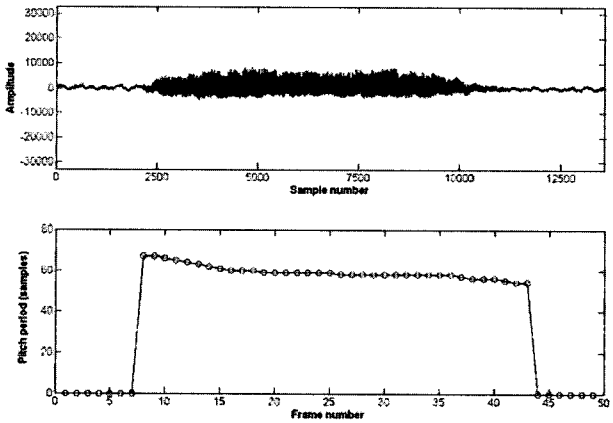
로그를 취한 필터뱅크 값에 IDCT (Inverse Discrete Cosine Transform)을 하여 최종 MFCC를 구한다.

나. 피치

음성 신호의 기본 주파수로 정의되는 피치정보는 음성을 이용한 화자인증 분야에서 널리 사용되고 있는 특징 파라미터이다. 음성신호의 피치 검출 방법은 크게 시간영역과 주파수 영역에서 추출하는 방법으로 나눌 수 있다. 시간영역에서의 피치 검출 알고리즘은 파형의 주기성을 강조하여 피치를 검출하는 방법으로서 병렬처리법, ACF (autocorrelation function), AMDF (average magnitude difference function) 등이 있다. 반면 주파수 영역에서의 피치검출 알고리즘은 음성 스펙트럼의 고조파 간격을 측정하여 기본주파수를 검출하는 방법으로서 켑스트럼법, 고조파분석법 등이 있으며, 시간영역에서의 방법보다 많은 연산량이 필요하다. 시간영역에서의 피치 추출 알고리즘은 분석을 위한 영역의 변환이 불필요하며, 합과 차 그리고 비교논리 등과 같은 간단한 연산만을 사용하게 되어 처리속도 측면에서의 주파수 영역에서의 방법보다 우수한 특성을 가진다.



(a) 남성 화자에 대한 35dB의 음성과 피치 검출결과



(b) 여성 화자에 대한 15dB의 음성과 피치 검출결과

그림 6. 발성음 "이"에 대한 음성파형과 피치검출 결과
Fig. 6. Waveform and pitch extraction result of voice /i/.

본 논문에서는 비교적 연산량이 작은 ACF 방법을 이용하여 피치를 추출하고, 음성 인증의 특징 파라미터로 이용하였다. 그림 6은 본 논문에서 획득한 음성데이터의 예시로서 (a)는 35dB, (b)는 15dB에 대한 음성 파형과 ACF방법을 이용하여 검출한 피치를 보여준다. 그림에 나타나듯이 발성음 /이/에 대한 음성 파형은 간단한 단모

음만을 포함하기 때문에 비교적 안정된 피치 정보를 갖는 장점이 있다.

2. GMM

입력 음성으로부터 추출한 MFCC와 피치정보는 GMM 알고리즘으로 모델링하여 음성인증에 적용된다. 이때 MFCC 계수 13차와 1차의 피치 계수는 파라미터 차원에서 결합되어, 최종적으로 새로운 14차의 특징 파라미터가 GMM의 입력 벡터로 사용된다. GMM은 주어진 데이터를 몇 개의 군집으로 나누고, 각 군집에 대한 가우시안 분포인 mixture component를 구한 후, 이들을 선형적으로 결합하여 하나의 가우시안 분포로 표현하는 방법이다. M개의 mixture component를 갖는 Gaussian mixture의 확률 밀도 함수는 식 (5)와 같이 표현된다[10].

$$p(x|\lambda) = \sum_{i=1}^{M} p_i b_i(x) \tag{5}$$

식 (5)에서 x는 D차원의 입력 데이터를 의미하여, $b_i(x)$는 i번째 가우시안 분포를 p_i는 각각의 가우시안 분포에 대한 가중치를 의미한다. D차원의 입력 데이터에 대한 가우시안 분포는 식 (6)과 같이 표현되며, 본 논문에서 D는 14차를 사용한다.

$$b_i(x) = N(x, u_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp[-\frac{1}{2}(x-u_i)^T \Sigma_i^{-1}(x-u_i)] \tag{6}$$

식 (6)에서 u_i 는 평균벡터이며 Σ_i는 공분산 행렬을 나타낸다.

GMM의 학습은 위의 모델 파라미터를 추정하는 과정으로 EM (Expectation Maximization) 알고리즘을 사용하여 파라미터를 추정하였다. EM알고리즘은 초기모델 λ로부터 $p(x|\bar{\lambda}) \geq p(x|\lambda)$인 새로운 모델 $\bar{\lambda}$을 추정하고, 이러한 반복 과정을 임계값으로 수렴할 때까지 계속하는 방법이다. 음성인증은 MFCC와 피치 계수의 결합된 특징 파라미터($\tilde{x}$)를 사전에 생성한 모델 λ와 유사도 값, $p(\tilde{x}|\lambda)$을 계산하여 임계값과 비교하고, 사용자의 인증 여부를 판단하는 방식으로 수행된다.

Ⅳ. 멀티모달 화자인증 시스템

그림 7은 치열영상과 음성을 이용한 멀티모달 화자인증 시스템의 블록도를 나타낸다. 입력 영상은 Haar-like 특징 기반의 AdaBoost 알고리즘이 적용되어 치열영역을 검출한다. 검출된 치열 영상에서 2D-DCT 특징벡터가 추출되고, EHMM 알고리즘을 이용하여 사용자의 치열영상에 대한 유사도가 계산된다. 또한 입력 음성은 전처리 과정이 수행되고, 전처리된 음성을 이용하여 MFCC와 피치 파라미터를 추출한다. MFCC와 피치 파라미터는 특징 단계에서 결합되고, GMM 알고리즘에 의하여 사용자의 음성에 대한 유사도를 계산한다. 전체적인 사용자 인증 성능을 높이기 위하여 본 논문에서는 치열영상과 음성에 대한 유사도 값들을 가중치의 합으로 결합하였다. 두 개의 단일 생체로부터 계산되는 유사도 값들은 그 분포 범위와 의미가 서로 다르기 때문에 0부터 1사이의 값으로 정규화되는 과정이 필요하다.

1. 유사도 정규화

두 가지의 단일 생체인증 시스템이 결합되었을 경우에

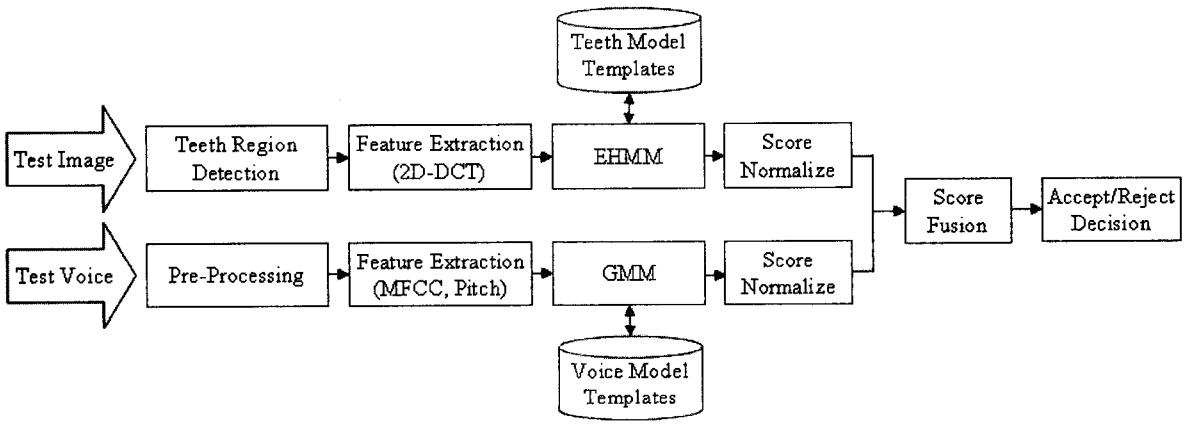


그림 7. 치열과 음성을 이용한 멀티모달 화자인증 시스템
Fig. 7. Multimodal speaker verification system using teeth and voice.

하나의 시스템은 데이터간의 거리를 도출하고 다른 시스템은 데이터간의 유사도를 도출한다면, 두 시스템에서 나온 정보를 결합하는 것은 무의미하며 잘못된 결과를 도출하게 된다. 또한, 여러 시스템으로부터 나온 유사도가 동일한 범위의 유사도를 도출하지 않기 때문에 가공되지 않은 상태로는 결합될 수 없으며, 각각의 독립적인 생체 정보에 의해 산출된 거리 또는 유사도와 같은 결과 값은 그 분포 범위와 의미가 서로 다르기 때문에 정규화 과정을 수행해야 한다. 일반적으로 널리 사용되고 있는 정규화 방법에는 최소-최대 정규화, Z-Score를 이용한 정규화, 10진수 변환기법 정규화, sigmoid 함수를 이용한 정규화 등의 다양한 방법이 있다. 본 논문에서는 치열과 음성으로부터 얻은 유사도를 0부터 1사이의 범위로 정규화하기 위하여 sigmoid 함수를 이용한 정규화 방법을 사용하였다[11]. 식 (7)과 (8)은 sigmoid 함수를 이용한 정규화 방법을 나타낸다.

$$o_i = \frac{1}{1+\exp(-\tau_i(o_{i,orig}))} \tag{7}$$

$$\tau_i(o_{i,orig}) = \frac{o_{i,orig} - (\mu_i - 2\sigma_i)}{2\sigma_i} \tag{8}$$

위 식에서 $o_{i,orig}$는 사용자의 개별 인증시스템(i)에서 얻은 유사도, o_i는 정규화한 유사도를 의미하며 μ_i와 σ_i는 각각 사용자의 개별 인증시스템(i)에 대한 평균값과 표준편차를 의미한다.

2. 치열과 음성정보의 결합

치열영상과 음성으로부터 계산되는 유사도 값들은 시스템의 전체적인 사용자 인증 성능을 높이기 위하여 결합된다. 처음의 과정은 각각의 단일 생체 인증 시스템으로부터 얻은 유사도 값을 정규화 하는 과정이며, 다음으로 정규화한 유사도 값들을 결합하여 단일한 값을 만들어 낸다. 두 개 이상의 단일 생체 인증 시스템의 유사도 값들을 결합하는 방법으로는 최대 유사도 선택, 최소 유사도 선택, 유사도의 가중치 합 등의 비교적 간단하면서도 높은 성능을 보이는 방법들이 있다. 최대 유사도 선택 방법은 단일 생체 정보로부터 얻은 유사도 값들 중에서 가장 큰 유사도 값을 선택하는 방법이며, 반면 최소 유사도 선택 방법은 가장 작은 유사도 값을 선택하는 방법이다. 이에 비해 유사도의 가중치 합을 이용한 방법은 단일 생체 정보로부터 얻은 유사도 값들에 각각 다른 가중치 값을 부여하여 새로운 유사도 값을 만드는 방법이다[12]. 본 논문에서는 두 개 이상의 단일 생체인식 시스템의 결합에 널리 사용되고 있는 가중치 합을 이용한 방법을 적용하여 치열과 음성에 대한 유사도 값들을 결합하였다. 가중치 합을 이용한 결합 방법은 이동단말 장치의 제한된 하드웨어적 리소스를 고려하여 선택되었으며, 간단한 결합 방식들 중에서 비교적 높은 성능을 보인다. 가중치 합을 이용한 결합 방법은 식 (9)와 같이 표현할 수 있다.

$$S_m = pS_t + (1-p)S_s, \quad 0 \le p \le 1 \tag{9}$$

식 (9)에서 S_m은 두 개의 유사도를 결합하여 생성한 단일의 유사도를 의미하고, S_t와 S_s는 각각 치열과 음성에 대한 유사도를 의미한다. 또한 p는 치열의 유사도에 부여하는 가중치이며, 음성의 유사도에 대한 가중치는 $(1-p)$로 모든 가중치의 합은 1이 되어야 한다.

V. 실험 및 결과

1. 실험 환경

본 논문의 멀티모달 화자인증 시스템은 HP iPAQ rw6100 기종의 스마트폰 환경에서 embedded Visual C++ 4.0의 프로그래밍 도구를 사용하여 구현하였다. 입력 영상은 스마트폰 장치에 내장된 카메라를 이용하여 480 640의 해상도로 획득되었고, 음성은 16 kHz의 샘플링율과 16bit/sample 품질로 획득하여 시스템의 성능평가를 수행하였다. 치열영상과 음성 데이터는 실제 환경에서 사용자가 움직이지 않는 상태를 전제로 획득

그림 8. 멀티모달 화자인증 시스템의 구현 화면
Fig. 8. Implemented display of multimodal speaker verification system.

되었다. 멀티모달 화자인증 시스템의 성능평가 실험에는 남자 20명과 여자 20명에 대하여 개인당 20개의 영상과 음성으로 구성된, 총 800개의 치열영상과 음성으로 구성된 데이터베이스를 이용하였다. 800개의 영상과 음성 데이터 중에서 40명에 대한 200개의 영상과 음성 데이터는 멀티모달 화자인증 시스템의 모델 학습에 이용하고, 나머지 600개의 영상과 음성 데이터는 성능평가 실험에 사용하였다. 그림 8은 데이터베이스 구축 환경의 예를 보이는 것으로 사용자는 움직이지 않는 상태를 전제로 하며, 주변 환경은 실내인 경우와 실외인 경우를 모두 포함하여 데이터베이스를 구축하였다.

2. 실험 결과

영상과 음성을 이용한 멀티모달 화자인증 시스템의 성능평가를 위하여, 본 논문에서는 치열영역의 검출 성능, 멀티모달 화자인증 시스템의 사용자 등록과 인증에 소요되는 시간, 그리고 시스템의 인증률에 대한 실험을 수행하였다. 표 1은 40명의 사용자에 대한 800개 영상에서 치열영역을 검출한 결과를 보이고 있다. 치열영역 검출률은 98.87%를 보였으며, 한 개의 영상에서 치열영역을 검출하는 평균 시간은 2.92초를 보였다.

표 2는 본 논문에서 구현한 멀티모달 화자인증 시스템의 사용자 등록과 인증에 소요되는 평균 시간을 보여준다. 사용자 등록에는 5개의 영상과 음성을 사용하여 모델을 생성하며, 사용자 인증에는 1개의 영상과 음성을 사용한다. 사용자 등록과 인증에 소요되는 평균 시간의 측정은 치열과 음성부분, 그리고 전체적으로 소요되는 시간을 구분하여 평가하였다. 실험 결과, 사용자 등록에는 평균 55.97초의 시간이 소요되었고, 사용자 인증에는 평균 10.76초가 소요되었다.

그림 9는 본 논문에서 구현한 멀티모달 화자인증 시스템의 구현 화면 예로서, 입력 영상에서 검출된 치열영상과 끝점이 검출된 음성을 이용하여 사용자를 인증하고 있는 상태를 보이고 있다. 그림 10은 본인과 사칭자에 대한 치열과 음성의 정규화한 유사도 값을 2차원 그래프로 도시한 결과를 보여준다. 이와 같은 유사도 값의 분포를 기반으로 본 논문에서는 유사도의 가중치의 변화에 따른 EER (Equal Error Rate)의 값을 조사하였으며, 실험에는 치열영역 검출에 성공한 영상과 이와 관련된 음성만을 사용하였다.

표 1. 치열영역 검출 실험결과
Table 1. Experimental results of teeth region detection.

구 분	AdaBoost 방법
검출률	98.87 %
오검출률	1.13 %
검출시간	2.92 s

표 2. 멀티모달 화자인증 시스템의 처리시간
Table 2. Processing time of multimodal speaker verification.

구 분	사용자 등록	사용자 인증
치열 인증	15.74 s	2.98 s
음성 인증	40.23 s	7.78 s
멀티모달 화자인증	55.97 s	10.76 s

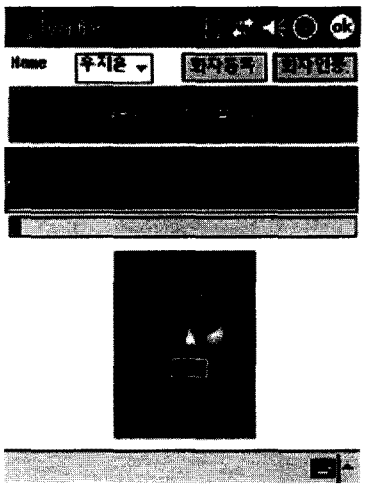


그림 9. 멀티모달 화자인증 시스템의 구현 화면
Fig. 9. Implemented display of multimodal speaker verification system.

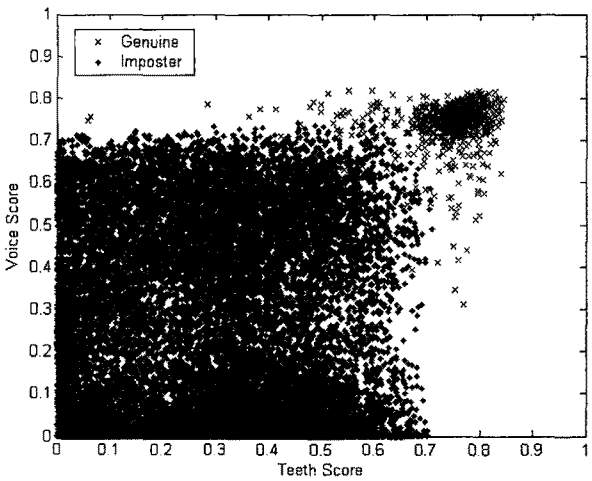


그림 10. 치열과 음성의 유사도 분포
Fig. 10. Distribution of teeth and voice score.

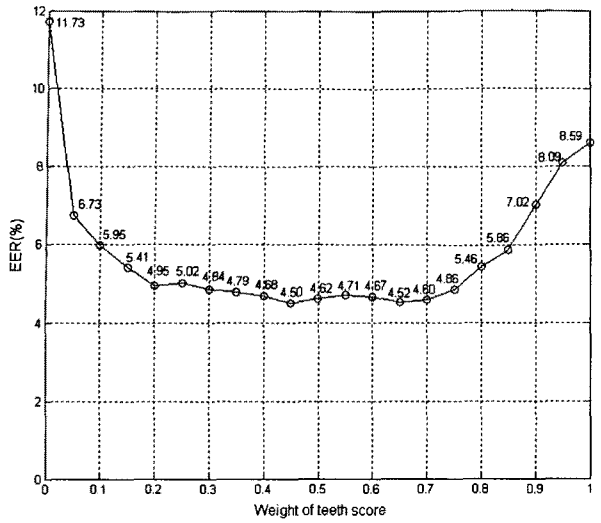


그림 11. 가중치에 따른 EER의 변화
Fig. 11. Result of EER by weight of teeth score

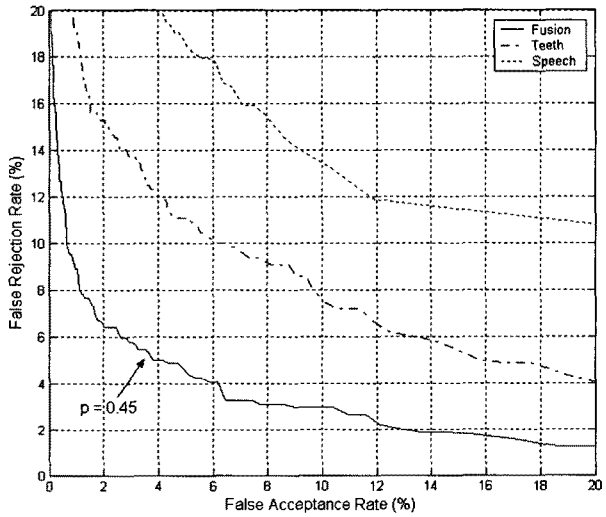


그림 12. 멀티모달 화자인증 시스템의 ROC 곡선
Fig. 12. ROC curves of multimodal speaker verification system.

그림 11은 치열의 가중치 p의 값을 0부터 1까지 0.05의 단계로 변화하여 계산한 멀티모달 화자인증 시스템의 EER를 나타낸다. 이 실험은 가중치 p의 값을 변화시켜 최적의 인증 성능을 얻기 위하여 수행되었다. 그림 11에서 단일 생체 인증 시스템의 EER는 치열인증의 경우 8.59%, 음성인증의 경우에는 11.73%를 보였다. 반면 가중치를 부여하여 합산된 유사도 값을 바탕으로 결과를 도출한 결과, p의 값이 0.45일 때, EER이 4.50%로 가장 좋은 성능을 보였다.

그림 12는 치열의 가중치 p가 0.45일 때, 치열과 음성, 그리고 두 개의 정보를 결합한 멀티모달 시스템의 성능을 ROC (Receiver Operating Characteristic) 곡선으로 나타내고 있다. 그림에서 가중치의 합으로 결합한 멀티모달 시스템의 경우는 치열 또는 음성만을 단독으로 사용한 경우보다 향상된 성능을 보였으며, 치열인증의 결과보다는 4.09%, 음성인증의 결과보다는 7.23%로 더 나은 성능을 보였다.

본 논문에서 제안한 방법은 이동단말 장치중의 하나인 스마트폰에서 구현하여 우수한 성능을 보임을 실험으로부터 확인하였다. 특히 일반 PC에서 획득한 고품질의 영상과 음성보다 스마트폰에서 획득한 영상과 음성이 저품질의 데이터임에도 불구하고, 치열인증과 음성인증의 결과는 비교적 좋은 성능을 보였다. 또한 이동단말 장치의 제한된 리소스를 고려하여 비교적 간단하면서도 우수한 성능을 보이는 가중치의 합으로 치열영상과 음성을 결합하여 보다 우수한 성능 향상을 얻을 수 있었다.

VI. 결 론

본 논문에서는 이동환경에서 개인의 신원을 인증하는 수단으로 치열영상과 음성을 생체정보로 이용한 멀티모달 화자인증 방법에 대하여 제안하였다. 제안한 방법은 이동단말 장치중의 하나인 스마트폰을 이용하여 치열인증과 음성인증 모듈로 구성하였으며, 치열인증 모듈은 2D-DCT 특징벡터와 EHMM 알고리즘으로 구성하고, 음성인증 모듈은 MFCC와 피치의 특징벡터를 GMM 알고리즘을 이용하여 사용자를 모델링하였다. 더불어, 제안한 방법은 이동단말 시스템의 제한된 리소스를 고려하여 치열과 음성의 유사도 결과를 비교적 간단하면서도 우수한 성능을 보이는 가중치의 합으로 결합함으로써, 전체적인 성능 향상을 도모하였다. 이를 위하여 40명에 대한 800개의 치열영상과 음성으로 성능평가 실험을 수행하였으며, 실험 결과, 치열인증과 음성인증의 결과는 각각 8.59%, 11.73%의 EER를 보였다. 또한 치열인증에 대한 가중치 p의 값을 0부터 1까지 0.05의 단계로 변화하여 인증 성능을 조사한 결과, p의 값이 0.45일 경우에 가장 좋은 4.05%의 멀티모달 인증 결과를 얻을 수 있었다. 실험 결과는 스마트폰 장치에서 획득한 치열영상과 음성이 범용 PC보다 저 품질임에도 불구하고 높은 인증 성능을 보였으며, 더불어 멀티모달 화자인증 시스템을 비교적 간단한 가중치 합으로 결합하여 각각의 단일 생체인증의 결과보다 향상된 인증 성능을 얻을 수 있었다.

본 논문에서 제안한 방법은 이동단말 장치에 대한 보안뿐만 아니라 기존에 사용하던 생체인식 기술을 대체하거나 또는 더불어 사용될 수 있음을 실험으로부터 확인하였고, 이동단말 환경에서 추가적인 장치의 부가 없이

활용될 수 있으므로 매우 경제적이고 신뢰성 있는 기술로 사료된다. 본 논문에서 제안한 방법은 스마트폰과 같은 임베디드 환경에서 구현되어 사용자 등록과 인증에 많은 시간이 요구되므로, 향후에 알고리즘 개선 및 작성 코드의 최적화와 같은 과정을 통하여 처리시간의 향상에 대한 연구가 필요하다.

참 고 문 헌

[1] A. K. Jain, A. Ross, and Prabbakar, "An introduction to biometric recognition", IEEE Trans. Circuits System, Video Technology, vol.14, no.1, pp.4-20, Jan. 2004.

[2] E. C. Epp, "Relationship Management: Secure Collaboration in a Ubiquitous Environment", IEEE Pervasive Computing, Volume 2, Issue 2, April , Pages 62-71, 2003.

[3] 권만준, 양동화, 고현주, 김진환, 전명근, "PDA를 이용한 실시간 얼굴 인식 시스템 구현", 퍼지 및 지능시스템학회 논문지, Vol. 15, No. 5, pp. 649-654, 2005.

[4] Tae-Woo KIM and Tae-Kyung CHO, "Teeth Image Recognition for Biometrics", IEICE TRANSACTIONS on Information and Systems Vol. E89-D No. 3 pp. 1309-1313, 2006.

[5] K. Prajuabklang, P. Kumhom, T. Maneewarn, and K. Chamnongthai, "Real-time Personal Identification from Teeth-image using Modified PCA", Proceeding, the 4-th information and computer Engineering Postgraduate Workshop, Vol. 4, No. 1, pp.172-175, 2004.

[6] C. Nadee, P. Kumhom, and K. Chamnongthai, "Improved PCA-Based Personal Identification Method Using Invariance Moment", The third International Conference on Intelligent Sensing and Information Processing, December 14-17, 2005.

[7] Dong-Ju Kim, Jong-Bae Jeon and Kwang-Seok Hong, "Performance Evaluation of Feature Vectors for Teeth Image Recognition", The 4th Conference On New Exploratory Technologies, October 25-27, 2007.

[8] P. Viola and M. J. Jones, "Robust real-time object detection", Technical Report Series, Compaq Cambridge research Laboratory, CRL 2001/01, Feb. 2001.

[9] A. V. Nefien and M. H. Hayes, "An embedded HMM-based approach for face detection and recognition", In Proc, IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 6, pp. 3553-3556, 1999.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data wia the EM algorithm", Journal of the Royal Statistical Society B, 1977.

[11] A. Ross and A. K. Jain, "Information fusion in biometrics", Pattern Recognition. Letter. 24 (13) 2003.

[12] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information", Digital Signal Processing, Volume 14, Issue 5, September Pages 449-480, 2004.

저 자 소 개

김 동 주(학생회원)

1998년 충북대학교 전파공학과 학사 졸업.
2000년 충북대학교 전파공학과 석사 졸업.
2008년 현재 성균관대학교 정보통신공학부 박사 과정.
<주관심분야 : 생체인식, 음성인식, 영상인식>

홍 광 석(정회원)

1985년 성균관대학교 전자공학과 학사 졸업.
1988년 성균관대학교 전자공학과 석사 졸업.
1992년 성균관대학교 전자공학과 박사 졸업.
1990년～1993년 서울보건전문대학 전산정보처리과 전임강사.
1993년～1995년 제주대학교 정보공학과 전임강사.
1995년～현재 성균관대학교 정보통신공학부 교수.
<주관심분야 : 신호처리, HCI, 음성인식, 영상인식>

하 길 람(학생회원)

2007년 세명대학교 전자공학과 학사 졸업.
2008년 현재 성균관대학교 전자공학과 석사 과정.
<주관심분야 : 영상인식, HCI>