MCE 학습 알고리즘을 이용한 문장독립형 화자식별의 성능 개선*

김태진(대전대), 최재길(대전대), 권철홍(대전대)

<차 례>

<Abstract>

Performance Improvement of a Text-Independent Speaker Identification System Using MCE Training

Tae-Jin Kim, Jae-Gil Choi, Chul-Hong Kwon

In this paper we use a training algorithm, MCE (Minimum Classification Error), to improve the performance of a text-independent speaker identification system. The MCE training scheme takes account of possible competing speaker hypotheses and tries to reduce the probability of incorrect hypotheses. Experiments performed on a small set speaker identification task show that the discriminant training method using MCE can reduce identification errors by up to 54% over a baseline system trained using Bayesian adaptation to derive GMM (Gaussian Mixture Models) speaker models from a UBM (Universal Background Model).

* Keywords: Speaker identification, MCE, GMM, UBM.

* 본 연구는 산업자원부의 지역혁신 인력양성 사업의 지원으로 수행되었음.

1. 서 론

음성을 이용한 통신의 기본적인 목적은 메시지 전달에 있다. 하지만 음성은 메시지뿐만 아니라 개인의 신원이나 구사된 언어의 종류, 화자의 심리적, 육체적, 감정상의 상태에 대한 정보를 포함하고 있다. 이러한 정보를 이용한 화자의 신원 파악, 즉 화자인식이 최근 관심의 대상이 되고 있다. 음성은 가장 쉽고 편리한 인간-기계 인터페이스로, 이를 이용한 화자인식 기술은 카드나 열쇠 보다 편리하고, 분실위험이 전혀 없어 안전하며, 손이나 다른 도구를 필요로 하지 않으므로 유비쿼터스 정보화시대의 중요한 인터페이스 기술로 자리매김하고 있다.

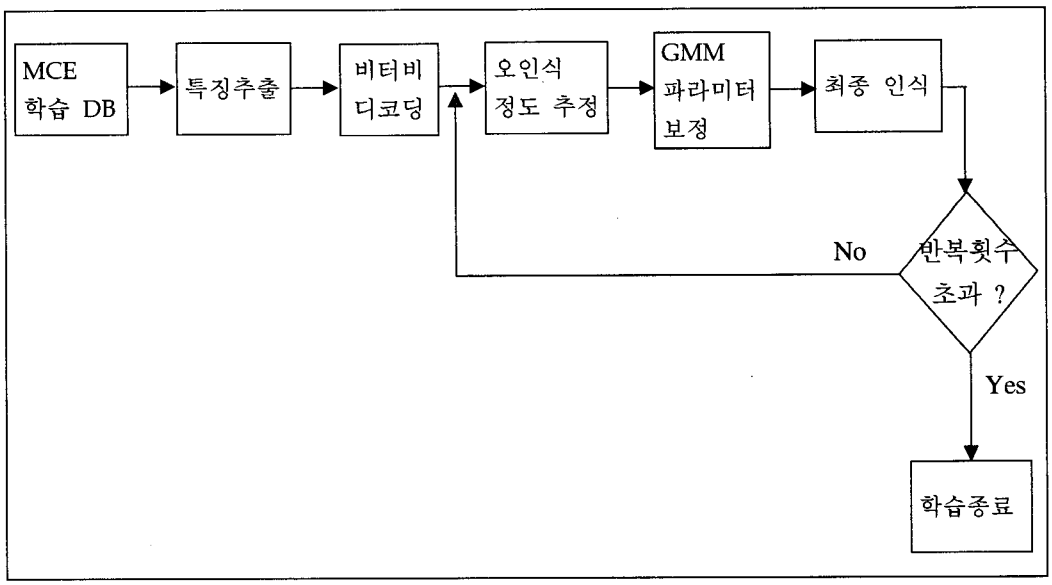
화자인식은 음성의 특징이 화자간의 변이가 화자 내의 변이에 비해 상대적으로 크다는 성질을 이용하여 화자를 구분하는 것으로 화자인증과 화자식별로 나뉜다. 본 논문에서는 화자식별을 다루는데, 이는 신원을 청구한 화자의 음성을 입력으로 하여 등록된 화자의 모든 성문모델과 비교하여 유사도가 가장 높은 화자를 선택하는 시스템이다. 화자인식은 인식대상이 되는 음성의 발성 방법에 따라 문장종속형과 문장독립형으로 나뉜다[1]. 화자가 발성할 문장이 정해져 있는 경우 문장종속형이라 하고, 문장이 정해져 있지 않고 자유롭게 발성하는 경우를 문장독립형이라 한다. 문장종속형은 발성어휘가 유출되었을 경우, 시스템의 보안에 치명적인 영향을 주게 된다. 문장독립형은 인식을 위해 사용하는 어휘를 임의로 자유롭게 발성하므로 문장종속형 보다는 녹취 등의 유출에 더 효과적으로 대처할 수 있다. 본 논문에서는 문장독립형 화자식별을 다룬다.

화자식별에서 고려해야할 사항은 적절한 인식모델 생성방식 및 특징 파라미터의 선택이다. 현재 대부분의 음성인식 시스템은 모델 생성을 위해 음성의 시간적 변화를 통계적으로 모델링할 수 있는 HMM(Hidden Markov Models)을 사용하고 있으나, 화자식별 분야에서와 같이 학습 데이터가 적은 경우에 인식 성능이 급격히 저하되는 단점이 있다. 최근 기계학습에 관한 연구가 활발히 진행되면서 MCE(Minimum Classification Error), SVM(Support Vector Machine) 등은 적은 학습 데이터에 대해서도 성능이 뛰어나다고 알려져 있다.

본 논문에서는 MCE를 화자식별에 적용하여 인식 실험을 수행하고 기존 GMM(Gaussian Mixture Models) 방식과 성능을 비교하여 MCE 방식의 우수성을 확인하고자 한다. 본 논문의 구성은, 서론에 이어 2장에서 제안한 MCE 학습 방식을 설명하고, 3장에서 실험 방법 및 결과를 논하고, 그리고 4장에서 결론 및 향후 연구과제에 대하여 기술한다.

2. MCE 학습 알고리즘

MCE는 인식오류를 최소화하는 변별적 학습방법이다. 이는 인식오류확률의 최소값을 구하는 대신 인식오류에 의해 구해지는 비용함수를 최소화하는 방법이다. <그림 1>은 MCE 학습 방법의 전체 개요를 보여 준다. 수집한 음성 DB에서 특징계수를 추출하고 기존 MLE(Maximum Likelihood Estimation) 방식으로 생성된 GMM 화자모델을 이용하여 비터비 디코딩을 수행하여 인식결과를 구한다. 인식결과에서 오인식 정도를 추정하고 MCE 학습 방법을 이용하여 만들어진 새로운 파라미터 값으로 보정된 GMM 화자모델로 최종인식을 수행한다. 학습을 반복하는 경우에 인식률이 증가하다 감소하는데 인식률이 더 이상 증가하지 않으면 반복학습을 중단한다.



<그림 1> MCE 학습 방식

MCE 방식에서는 각 화자모델 Λ_i, $i = 1, ..., L$에 대한 변별함수(discriminant function) $g_i(X, \Lambda_i)$가 필요하다. L은 화자의 수, Λ_i는 화자별 GMM 파라미터 값을 나타낸다. 가장 큰 변별함수 값을 갖는 모델을 선택함으로써 입력 X가 결정된다. 여기에서는 비터비 디코딩에서 구한 log-likelihood 값이 변별함수로 사용된다.

본 논문에서는 화자 모델 GMM이 인식의 기본적인 단위로서 사용된다. j번째 상태에서 관측벡터 x의 출력확률 밀도함수는 식 (1)과 같다.

$$b_j(\boldsymbol{x}) = \sum_{m=1}^{M} c_{jm}\, N(\boldsymbol{x}, \mu_{jm}, \Sigma_{jm}) \tag{1}$$

여기에서 M은 상태의 mixture 개수이다. c_{jm}은 mixture weight이고 $\sum_{m=1}^{M} c_{jm} = 1$을 만족한다. $N(\boldsymbol{x}, \mu_{jm}, \Sigma_{jm})$은 평균 벡터 μ_{jm}과 대각선의 공분산 행렬 Σ_{jm}을 갖는 Gaussian 분포를 나타낸다.

i번째 화자모델 Λ_i에서 최적의 경로에 따른 입력 벡터열 $\boldsymbol{X} = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_T\}$의 log-likelihood 값 즉 변별함수는 식 (2)와 같이 나타낼 수 있다.

$$g_i(\boldsymbol{X}, \Lambda_i) = logb^i_{s_0}(\boldsymbol{x}_0) + \sum_{t=1}^{T} loga^i_{s_{t-1}s_t} + \sum_{t=1}^{T} logb^i_{s_t}(\boldsymbol{x}_t) \tag{2}$$

$S = \{s_0, ..., s_T\}$는 최적경로에 따른 상태 열을 나타낸다. T는 입력음성 $\boldsymbol{X}$의 프레임 수이다. $\log b^i_{s_t}(\boldsymbol{x}_t)$는 i번째 모델의 상태 s_t에서 관측벡터 $\boldsymbol{x}_t$의 출력확률의 로그 값이다. $a^i_{s_{t-1}s_t}$는 i번째 모델에서 상태 s_{t-1}에서 s_t로의 천이확률을 나타낸다. $c = \arg\max g_i(\boldsymbol{X}, \Lambda_i), i = 1, ..., L$를 만족하면, 인식기는 입력 $\boldsymbol{X}$를 c번째 화자의 음성으로 인식한다.

MCE에서 최소화 하고자 하는 최종 함수는 화자 식별에서의 오류 개수이다. 식 (3)은 오인식 정도를 나타내는 함수이다.

$$d_c(\boldsymbol{X}) = -g_c(\boldsymbol{X}, \Lambda_c) + \left[\frac{1}{L-1} \sum_{i \neq c} g_i(\boldsymbol{X}, \Lambda_i)^{\eta} \right]^{\frac{1}{\eta}} \tag{3}$$

여기서 η은 양수이고, L은 전체 화자의 수, $\boldsymbol{X}$는 c번째 화자에 대한 입력 음성이다. 우측 첫째 항은 인식대상 화자의 음성에 대한 log-likelihood 값이고, 둘째 항은 다른 화자들에 대한 log-likelihood 값의 합을 나타낸다. 이 식에서 큰 양수 값 $d_c(\boldsymbol{X})$은 오인식이 발생했다는 것을 의미한다. 또한, $d_c(\boldsymbol{X})$ 값의 부호와 절대값은 인식 오류가 발생 했는가 아닌가를 보여 주므로 이 식은 오인식 정도를 나타내는 함수라고 말할 수 있다.

다음으로, 인식오류의 최소화에 대해 비용함수를 정의한다. 식 (4)는 모델 i에 대한 비용함수이다.

$$l_i(d_i(\boldsymbol{X})) = \frac{1}{1+\exp(-\gamma(d_i(\boldsymbol{X}))} \tag{4}$$

여기에서 γ값은 고정된 값으로 비용함수의 기울기를 조절한다. 이 비용함수의 최소화는 오류 개수의 최소화로 직접 연결된다.

GPD(Generalized Probabilistic Descent) 방식이 효과적으로 GMM 파라미터 값을 반복적으로 조절하기 위해 사용된다[2]. 이 방법에서 입력 음성 $\boldsymbol{X}$에 대하여 화자 모델 파라미터 Λ는 다음 식 (5)에 의해 보정된다.

$$\Lambda_{n+1} = \Lambda_n - \epsilon_n U \nabla l_c(d_c(\boldsymbol{X})) \tag{5}$$

여기서 Λ_n은 현재 파라미터 값, Λ_{n+1}은 보정된 파라미터 값이고, 오른쪽 두 번째 항은 보정 정도를 나타낸다. U는 양수의 한정된 행렬이고, $\{\epsilon_n\}$은 $\sum_{n=1}^{\infty}\epsilon_n \rightarrow \infty$, $\sum_{n=1}^{\infty}\epsilon_n^2 < \infty$을 만족한다. 여기에서 mixture weight는 $\sum_m c_{jm} = 1$, $c_{jm} \geq 0$를, 분산은 양수 값을 가져야 한다는 조건을 만족해야 한다.

GMM 파라미터를 조절하기 위한 이 기울기(gradient) 방법에서 식 (5)의 오른쪽 두 번째 항을 구하는 방법은 다음과 같다. 우선 파라미터들의 대수를 구하고 gradient descent에 의해 값을 조절한다. 그 다음에 조절된 파라미터들의 멱지수를 구하고 마지막으로 위 조건을 만족시키기 위해 값을 정규화 한다[3]. 인식대상 화자 c에 대해 입력음성 화자 p의 GMM 파라미터에 대한 $l_c(d_c(\boldsymbol{X}))$의 미분계수 $(\nabla l_c(d_c(\boldsymbol{X})))$는 식 (6), (7), (8)과 같이 구할 수 있다.

$$\frac{\vartheta l_c(d_c(\boldsymbol{X}))}{\vartheta c_{jm}^p} = k \cdot \gamma_{jm}^p \cdot \left(-\sum_{t=1}^{T} sgn(c_{jm}^p, c_{s_t m}^c) + \frac{1}{M-1}\sum_{i\neq c}\sum_{t=1}^{T} sgn(c_{jm}^p, c_{s_t m}^i)\right) \tag{6}$$

$$\frac{\vartheta l_c(d_c(\boldsymbol{X}))}{\vartheta \sigma_{jmd}^p} = k \cdot \gamma_{jm}^p \cdot \left(-\sum_{t=1}^{T} sgn(\sigma_{jmd}^p, \sigma_{s_t md}^c)\right.$$
$$\left. + \frac{1}{M-1}\sum_{i\neq c}\sum_{t=1}^{T} sgn(\sigma_{jmd}^p, \sigma_{s_t md}^i)\right) \cdot \left(-\frac{1}{2} + \frac{1}{2}\cdot\frac{(x_{td}-\mu_{jmd}^p)^2}{\sigma_{jmd}^p}\right) \tag{7}$$

$$\frac{\vartheta l_c(d_c(\boldsymbol{X}))}{\vartheta \mu^p_{jmd}} = k \cdot \gamma^p_{jm} \cdot (- \sum_{t=1}^{T} sgn(\mu^p_{jmd}, \mu^c_{imd}) + \frac{1}{M-1} \sum_{i \neq ct=1}^{T} sgn(\mu^p_{jmd}, \mu^i_{jmd})) \cdot (\frac{x_{td} - \mu^p_{jmd}}{\sigma^p_{jmd}}) \quad (8)$$

위 식에서 $sgn(A,B)$는 A,B가 같은 모델의 파라미터에 속하면 1로, 그렇지 않으면 0으로 정의된다. x_{td}는 $\boldsymbol{x_t}$의 d번째 값이고, σ^p_{jmd}는 $\sum_{jm}^{p}$의 d번째 대각선 값이다. 그리고 k와 γ^p_{jm}는 다음과 같다.

$$k = \frac{\exp^{\gamma \cdot d_c} \cdot \gamma}{(1 + exp^{-\gamma \cdot d_c})^2} \quad (9)$$

$$\gamma^p_{jm} = \frac{c^p_{jm} N(\boldsymbol{x_t}, \boldsymbol{\mu^p_{jm}}, \sum_{jm}^{p})}{b^p_j(\boldsymbol{x_t})} \quad (10)$$

3. 실험 방법 및 결과

3.1 음성 DB

실험에 사용한 음성 DB는 ETRI의 음성정보연구센터에서 구축한 비영리 목적의 화자인식용 DB를 사용하였다. 본 DB는 사무실 PC 환경에서 중가의 PC 마이크(모델명: Shenheiser MD425)를 이용하여 남자 25명, 여자 25명 등 총 50명의 화자가 발성한 2연 숫자, 4연 숫자, 문장으로 구성되어 있다. 이 중에서 본 논문에서 실험에 사용한 것은 문장음성으로, 문장의 발성목록은 개인정보와 관련된 10개의 단어와 3어절 이내로 구성된 단문 10개 등 20문장으로 구성되어 있고, 한 화자당 한 차수에 동일한 목록을 5회 발성하고, 주차 간격으로 20명의 화자가, 월차 간격으로 다른 20명의 화자가, 3개월차 간격으로 다른 10명의 화자가 20문장을 각각 4회 반복하여 녹음 수집한 것이다. 본 논문에서는 훈련용으로 50명의 화자가 발성한 10개의 단어를 사용하였고, 테스트용으로 3어절 이내로 구성된 10개의 단문을 사용하여 문장 독립형 화자식별의 성능을 실험하였다.

음성신호를 매 10msec 마다 25msec의 Hamming 창함수를 사용하여 분석하였는

데, 음성의 특징 파라미터로 MFCC(Mel Frequency Cepstrum Coefficients)를 사용하여 12차 MFCC, delta coefficients, acceleration coefficients, 에너지, delta 에너지, acceleration 에너지 등 총 39차를 추출하였고, HMM의 구조는 1-state GMM으로 화자별 음향모델을 생성하였다.

3.2. UBM의 화자적응을 통한 화자식별

화자 식별 실험의 통일성 및 신뢰성을 위해서 본 논문에서는 mixture 수를 결정하기 위한 실험을 하였다. 남녀 50명에 대해 화자별로 GMM 모델(앞으로는 이 방식을 화자별 GMM이라고 언급한다.)을 만들고 mixture 수를 2, 4, 8, 16, 32, 64, 128, 256으로 늘리면서 화자 식별 실험을 하였다. 화자별 GMM 모델은 화자별로 10개의 단어를 사용하여 만들었다. 남녀 50명의 화자가 발성한 10개의 단문으로 화자 식별한 결과는 <표 1>과 같다. <표 1>을 보면, 실험 결과 mixture 수가 64일 때 인식률이 92.2%로 가장 좋은 결과를 보였다. 따라서 본 논문에서의 모든 실험은 mixture 수를 64로 하였다. <표 1>에서 mixture 수가 256 이상 증가하면 인식률이 급격히 감소하는데, 이는 학습단어 수가 적기 때문에 나타난 결과라고 생각된다.

<표 1> mixture 수에 따른 화자식별 실험결과

mixture 수	2	4	8	16	32	64	128	256
인식률(%)	65.8	74.4	85.4	90.4	91.8	92.2	91.6	79.0

화자식별 분야에서 성능이 좋다고 알려져 있는 시스템은[4] UBM(Universal Background Model)이라는 배경화자 모델을 사용하고, 이 모델에 화자적응을 적용하여 각 화자의 모델을 생성하는 방식으로(앞으로는 이 방식을 UBM-adapted GMM이라고 언급함), 본 논문에서는 이를 성능비교 대상으로 삼았다.

본 논문에서 UBM을 생성한 방법은 다음과 같다[5]. 먼저 성별로 각각 모델을 생성한 다음 이들을 조합하여 하나의 UBM을 만들었다. 우선 남녀 각각 25명의 DB를 이용하여 mixture 32를 갖는 남녀 UBM을 만들었다. 그리고 각각의 남녀 UBM 모델에서 mixture weight를 정규화 과정을 통하여 수정한 후 두 모델을 합하여 하나의 UBM을 만들었다. 두 모델을 단순히 합하면 mixture weight의 합이 2가 되므로 정규화를 통하여 그 합이 1이 되도록 조정하였다. 이 통합된 UBM의 mixture 수는 64가 된다. 다음에는 UBM으로부터 각 화자의 음성 DB를 이용하여 적응방법을 통하여 각각의 화자모델을 생성한다. 이 방식은 적은 학습 자료로 높은 인식 성능을 얻기 위한 방법이다. 앞에서 생성한 UBM을 이용하여 적응 과정

을 통해 각각의 화자모델에 대한 mean을 갱신하는 과정은 다음과 같다. 먼저 mixture i에 대하여 다음의 likelihood를 구한다.

$$\Pr(i|\boldsymbol{x}_t) = \frac{w_i p_i(\boldsymbol{x}_t)}{\sum_{j=1}^{M} w_j p_j(\boldsymbol{x}_t)} \tag{11}$$

$p_i(\boldsymbol{x}_t)$는 Gaussian 분포, w_i는 mixture weight, M은 mixture 수를 나타낸다. 그리고 다음을 차례로 구하여 UBM의 mean(μ_i)으로부터 화자모델의 mean($\hat{\mu_i}$)을 갱신한다.

$$n_i = \sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t) \tag{12}$$

$$E_i(\boldsymbol{x}) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|\boldsymbol{x}_t)\boldsymbol{x}_t \tag{13}$$

$$\hat{\mu_i} = \frac{n_i}{n_i + r} E_i(\boldsymbol{x}) + \frac{r}{n_i + r}\mu_i \tag{14}$$

여기에서 r은 실험값으로 16을 사용하였다. 본 논문에서는 mean, variance, mixture weight 중에서 mean 만을 갱신하여 화자모델을 생성하였다[5].

UBM-adapted GMM 방식에서, UBM을 만들기 위해 50명의 화자가 최초 발성한 10개의 단어 등 총 500개의 단어를, 화자적응을 위해서는 화자별로 10개의 단어를 사용하였다. 남녀 50명의 화자가 발성한 10개의 단문으로 화자 식별한 결과는 <표 2>와 같다. 화자별 GMM 방식의 오류율은 7.8%이고 UBM-adapted GMM 방식은 5.2%로 줄어 33% 정도의 상대 오류율 감소를 보여 준다.

<표 2> 화자별 GMM 방식과 UBM-adapted GMM 방식의 화자식별 결과 비교

	화자별 GMM	UBM-adapted GMM
오류 개수 / 총 테스트 문장 수	39 / 500	26 / 500
오류율 (%)	7.8	5.2

3.3. MCE를 적용한 화자식별

UBM-adapted GMM 방식으로 생성한 화자 모델에 MCE를 적용하여 화자 모델을 갱신하였다. MCE 학습에 사용된 음성 DB는, 단어수를 달리하면서 실험한 결과 화자별로 20개 단어(10개의 단어 X 2회 발성)에서 가장 좋은 성능을 보였다. <표 3>을 보면, MCE를 적용한 방식의 오류율이 2.4%이고 UBM-adapted GMM 방식이 5.2%로 상대 오류율이 54% 정도 개선됐음을 알 수 있다.

<표 3> UBM-adapted GMM 방식과 MCE를 적용한 방식의 화자식별 결과 비교

	UBM-adapted GMM	MCE 적용
오류 개수 / 총 테스트 문장 수	26 / 500	12 / 500
오류율 (%)	5.2	2.4

4. 결 론

본 논문에서는 문장독립형 화자식별의 성능 개선을 위하여 MCE 학습 방법을 제안하였다. MCE 학습 방법은 인식오류를 최소화하는 학습방법으로 이는 최소 인식오류 확률 값을 정확히 구하는 대신 인식 오류에 발생하는 비용함수를 최소화하는 방법으로, 본 논문에서는 이 방법이 화자식별에 유효한 방법인가를 검증하고자 하였다. 화자식별 분야에서 기존에 성능이 좋다고 알려져 있는 시스템인, UBM이라는 배경화자 모델을 만들고 이 모델에 화자적응을 통하여 각 화자의 모델을 생성하는 방식과 성능을 비교하였다. 실험 결과를 보면 MCE를 적용한 방식이 UBM-adapted GMM 방식 보다 화자식별 오류율이 54% 정도 감소됐음을 알 수 있다. 따라서 본 논문에서는 화자식별에서 MCE 알고리즘과 같은 변별적인 학습 방법의 우수성을 입증하였다.

향후 연구에서는 화자식별의 성능 개선을 위하여 다른 기계학습 기법인 SVM을 실험할 예정이다.

참 고 문 헌

[1] 대한음성학회, "음성정보기술(SIT) 로드맵", *대한음성학회 보고서*, 2003.

[2] W. Chou, B. H. Juang, C. H. Lee, "Segmental GPD training of HMM based speech recognizer", *Proc. ICASSP 92*, pp. 473-476, San Francisco, 1992.

[3] Y. J. Chung, C. K. Un, "Multilayer perceptrons for state-dependent weightings of HMM likelihoods", *Speech Communication*, Vol. 18, pp. 79-89, 1996.
[4] 유하진, "화자인식 기술 및 국내외시장 동향", *대한음성학회 2004 봄 학술대회 발표논문집*, pp. 91-97, 2004.
[5] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, "Speaker verification using adapted Gaussian mixture models,", *Digital Signal Processing*, Vol. 10, pp. 19-41, 2000.

접수일자: 2006년 2월 15일
게재일자: 2006년 3월 16일

▶ 김태진(Tae-Jin Kim)
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2567
E-mail: tjyh1004@daum.net

▶ 최재길(Jae-Gil Choi)
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2567
E-mail: u2u2u2u2u2@nate.com

▶ 권철홍(Chul-Hong Kwon) : 교신저자
주소: 300-716 대전광역시 동구 용운동 96-3 대전대학교
소속: 대전대학교 정보통신공학과 BMW 연구실
전화: 042) 280-2555
E-mail: chkwon@dju.ac.kr