# 모의 지능로봇에서의 음성 감정인식*

장광동(충북대), 김남(충북대), 권오욱(충북대)

## <차 례>



## <Abstract>

### Speech Emotion Recognition on a Simulated Intelligent Robot

Kwang-Dong Jang, Nam Kim, Oh-Wook Kwon

We propose a speech emotion recognition method for affective human-robot interface. In the proposed method, emotion is classified into 6 classes: Angry, bored, happy, neutral, sad and surprised. Features for an input utterance are extracted from statistics of phonetic and prosodic information. Phonetic information includes log energy, shimmer, formant frequencies, and Teager energy; prosodic information includes pitch, jitter, duration, and rate of speech. Finally a pattern classifier based on Gaussian support vector machines decides the emotion class of the utterance. We record speech commands and dialogs uttered at 2m away from microphones in 5 different directions. Experimental results show that the proposed method yields 48% classification accuracy while human classifiers give 71% accuracy.

*Keywords: Emotion recognition, Support vector machine, Speech interface.

* 이 논문은 2005년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음.

# 1. 서 론

음성은 사람들 사이에 의사소통을 하는데 있어 의미뿐만 아니라, 감정도 전달한다. 음성에 내포된 감정은 단어를 강조하거나 화자의 심리상태를 나타내어 의사소통을 더 자연스럽게 한다. 정서적 휴먼컴퓨터 인터페이스(affective human computer interface)는 최근 들어 휴머노이드형 로봇의 관심에 힘입어 많은 관심의 대상이 되고 있다. 사람의 감정을 인식하는데 있어 영상을 이용한 얼굴의 감정표현 인식과 음성을 이용한 감정인식 이용한 연구가 많이 되고 있다. 영상을 이용한 경우는 사람의 얼굴 표정에서 주요 특징인 입술, 눈, 코의 위치를 찾고 모양과 감정간의 기하학적인 관계를 파악하여 감정을 인식하는 방법이 시도되었다.

음성신호는 사전적인 범주와 운율적인 범주로 나누어 볼 수 있다. 사전적인 범주는 사람들이 서로 이해할 수 있는 단어들이며, 각 언어권에서 사용하는 각각의 발화들로 구성되어 있고, 운율적인 범주는 음성에 내포되어 있는 운율, 즉 음악적인 요소들로 구성되어 있는 것을 말한다. 음성에서의 운율적인 요소들은 청자로 하여금 화자의 감정을 예측할 수 있도록 하는 요소이다. 그러나 감정을 표현하는데 있어 일반적으로 감탄사를 말하는 경우가 있고, 일상적인 생활에서 사용되는 단어들에 감정이 표현된 경우가 있다. 이에 단어의 의미로부터 감정을 인식하는 방법, 단어의 의미와 상관없이 운율적인 정보만을 이용하는 방법, 그리고 두 가지를 모두 사용하는 방법 등에 대한 많은 연구가 있었다[3].

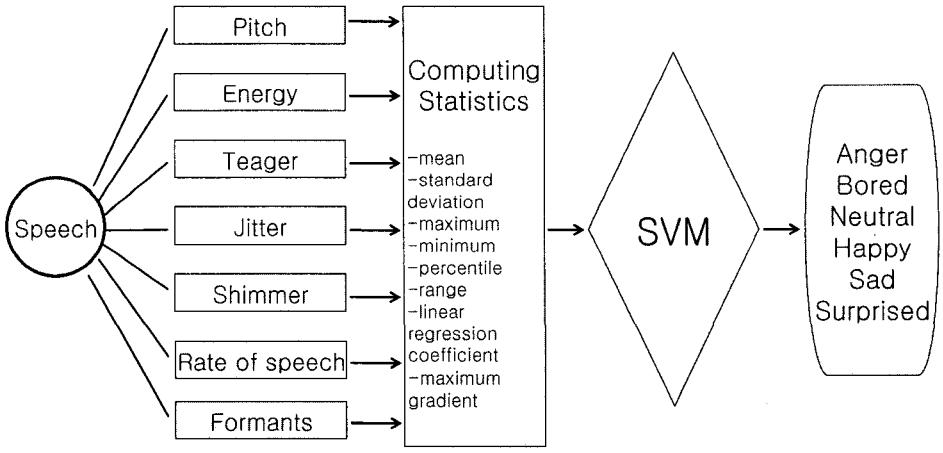
사람의 감정을 분류하는데 있어 여러 가지 분류기준이 있지만 대개 화남, 기쁨, 감정이 없는 상태, 슬픔, 놀람, 지루함, 혐오등과 같이 분류하고 있다. 그러나, 감정은 한순간에 하나의 감정으로만 표현되는 경우만 있는 것이 아니라, 복합적으로 한 개 이상의 감정이 동시에 나타나기도 한다. 가령, 기쁨과 놀람이 같이 나타날 수 있는 경우가 있다[1]. 감정을 인식하는데 있어 감정이 없는 상태와 감정이 있는 상태로 분류하여 감정인식 접근방법도 있다.

감정인식에 사용되는 특징들은 에너지, 포만트, 템포, 지속시간, 주파수변이(jitter), 진폭변이(shimmer), mel frequency cepstral coefficient (MFCC), linear predictive coding (LPC)계수, Teager에너지 등이 있다. 여러 특징들 중에 감정을 인식하는데 가장 큰 기여하는 특징은 피치와 에너지이다[2][4]. 추출된 특징들을 가지고 hidden Markov model (HMM)[5], support vector machine (SVM), neural network 등을 사용하여 감정을 분류한다.

본 연구의 목적은 음성에 내포되어 있는 음향정보와 운율 정보들로부터 유효한 특징들을 추출하여 지능 로봇이 이러한 정보를 이용하여 상황에 맞는 감정 상태를 인식함에 있다.

## 2. 감정인식

음성을 이용한 감정인식에서는 감정상태를 기쁨(happy), 슬픔(sad), 놀람(surprised), 지루함(bored), 화남(angry), 감정이 없는 상태(neutral)와 같이 6가지의 상태로 분류하였다. 음성으로부터 기본 특징들을 추출한 후 기본 특징으로부터 통계적인 값을 계산한 결과 값을 입력으로 한 SVM 패턴분류기를 사용하여 감정을 인식한다. 기본 추출 특징들로는 피치(pitch), 에너지, 주파수변이(jitter), 진폭변이(shimmer), Teager 에너지, 발화율(rate of speech (ROS)) 그리고 포만트(formant)를 사용하였고, 기본 특징들로부터 평균(mean), 표준편차(standard deviation), 최대(maximum), 최소(minimum), 퍼센타일(percentile), 범위(MAX-MIN), 선형회귀계수(linear regression coefficient), 최대 기울기(maximum gradient) 특징의 통계적인 값을 계산하여 사용하였다.




<그림 1> 음성 감정인식기

### 2.1. 특징 추출

16비트와 샘플링 주파수 16kHz인 음성신호는 노이즈를 제거하기 위해서 위너 필터링(Wiener filtering)한 후 끝점을 검출하여 음성부분만 추출하였다. 일반적으로 MFCC를 추출하여 음성인식을 하는 경우에는 해밍윈도우(Hamming window)를 사용하지만 이 논문에서는 추출된 음성을 윈도우 크기가 25ms인 해닝윈도우(Hanning window)를 사용하여 10ms단위로 구간 이동하면서 특징을 추출하였다.

### 2.1.1. 피치 추출

피치를 추출하는 방법으로는 우선 음성을 저 대역 필터링하여 10ms단위로 구간 이동하면서 프레임 크기가 60ms단위로 average magnitude difference function (AMDF)를 사용하여 계산한 피치 후보들에서 최소값을 피치로 결정한 후 smoothing하는 방법을 사용하였다. AMDF를 사용하여 피치를 추출하는 것은 피치를 정확히 계산하여 주는 방법은 아니지만 노이즈에 강한 특성을 보이고 계산이 빠르고 간단하다[8].

$$AMDF_n(j) = \frac{1}{N}\sum_{i=1}^{N}|x_n(i) - x_n(i+j)|,\ 1 \le j \le MAXLAG \tag{1}$$

여기서 $N$은 음성 샘플 개수, $x_n(i)$는 $n$번째 프레임의 $i$번째 음성 샘플 신호, $MAXLAG$은 피치 주기의 최대값이다.

### 2.1.2. 에너지

에너지는 일반적으로 많이 사용하는 로그에너지와 Teager 에너지를 사용하였다. Teager 에너지는 기존의 알고리듬[7]을 사용하여 추출하였다. Teager 에너지는 음성신호가 복합 정현신호로 구성되어 있으므로 각 주파수 대역으로 분류한 후 계산한다. 단일대역 주파수대로 분류하기 위해서 필터뱅크를 사용하여 주파수 대역별로 Teager 에너지를 구하였다.

$$E_{Teager}(n,i) = f_n^2(i) - f_{n+1}(i)f_{n-1}(i),\ i = 1...FB \tag{2}$$

여기서 $f_n(i)$ 프레임 $n$번째의 $i$번째 필터뱅크 계수, $FB$는 필터뱅크의 개수이다. Teager 에너지 추출 알고리듬의 장점은 세 개의 계수를 사용하여 계산하기 때문에 빠른 계산을 할 수 있다.

### 2.1.3. 주파수변이와 진폭변이

주파수변이(주파수변동률)와 진폭변이(진폭변동률)는 음성의 음질을 분석할 때 사용하는 특징으로서 주파수변이는 피치 주파수의 변동정도를 진폭변이는 피치 간에 진폭의 변동되는 정도를 나타낸다[9].

$$Jitt = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|T_{0_i} - T_{0_{i+1}}|}{\frac{1}{N}\sum_{i=1}^{N}|T_{0_i}|} \tag{3}$$

$$Shimm = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|A_i - A_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N}|A_i|} \tag{4}$$

여기서 $N$은 한 문장 또는 단어단위의 음성에서 AMDF를 사용하여 추출된 피치의 개수, $T_{0_i}$는 $i$번째의 피치 주기이고 $A_i$는 $i$번째 프레임의 진폭이다. 주파수변이와 진폭변이는 음질을 결정하는 요소로 유성음의 지속기간 동안 변화하는 정도를 알 수 있다. 주파수변이는 무성자음에 이어지는 유성음으로 변하는 구간에서 가장 높게 나타나고 분포가 넓게 되는 경우에는 음질이 거칠다고 한다. 진폭변이는 음성에서의 에너지 분포가 규칙 유무에 따라서 정상음성 또는 감정이 있는 음성인지를 알 수 있다.

### 2.1.4 발화율

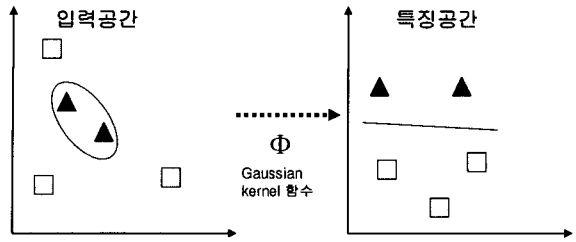
발화율은 음성을 유성음/무성음/무음(voice/unvoice/silence)구간으로 나누어 단위 구간의 모음비율을 나타낸 것이다. 발화율은 고정형과 가변형이 있는데 음성 감정 인식기에서는 고정 발화율를 사용하였다.

$$ROS = \frac{N}{\sum d_i} \tag{5}$$

여기서 $N$은 유성음 구간 개수이고, $d_i$는 $i$번째 유성음 구간의 지속시간이다.

## 2.2. 감정 분류

추출된 특징들로부터 평균, 퍼센타일, 표준편차, 최대, 최소, 범위, 선형회귀계수, 최대 기울기의 통계적인 값을 계산하여 leave-one-out 교차 검증을 적용하여 인식 훈련과 테스트에 사용하였다. SVM은 입력이 다차원인 경우에 사용하여 최적의 분류를 할 수 있는 방법으로 같은 부류에 속하는 데이터들을 같은 쪽에 위치하게 만드는 hyperplane을 찾으면서 부류간의 거리를 최대화하여 분류하는 방법이다 [10][11].




<그림 2> SVM의 개념

본 논문에서 사용한 SVM은 Gaussian kernel SVM 을 사용하여 감정을 인식하였다. 가우시안 커널(Gaussian kernel)의 개수는 10를 사용하였으며 감정을 훈련 및 테스트하는데 있어 6개의 감정 중 하나 이상의 감정이 동시에 인식되는 것은 배제하고 하나의 문장 또는 단어가 입력이 되면 하나의 감정만이 있는 것으로 가정하였다.

## 3. 실험결과

### 3.1. 음성 데이터베이스

감정 음성 데이터베이스는 지능형 로봇의 감정인식 인터페이스 개발을 위해 성우가 아닌 일반인으로 20~30대 30명(남녀 각 15명)의 화자로부터 6가지의 감정 -기쁨, 슬픔, 놀람, 지루함, 화남, 감정이 없는 상태를 녹음하였다.

마이크와 화자사이의 거리를 2m, 화자와 마이크 사이의 각도를 전방향 좌우 각도로 하여 스테레오로 녹음하였다. 무음구간은 300ms로 사용자 등록, 인사, 생활정보, 명령, 감정 등을 나타낼 수 있는 5개의 항목으로 구성되었으며, '사용자 등록'을 제외한 4개의 항목(50단어 및 문장)으로 6가지 감정을 발화하였다. 한 화자당 발화량은 302개이며 총 9,060개의 발화로 구성되어 있다.

<표 1> 감정 음성 데이터베이스 사용 단어

| 항목 | 단어 및 문장 |
|---|---|
| 사용자등록 | 사용자등록, 내 이름은 ○○○입니다. |
| 인사 | 안녕, 잘 지내어?, 보고 싶었어, 오랜만이야!, 뭐하고 놀았어?, 잘 있어, 갔다 올게, 나중에 봐, 빨리 와, 어디 가? |
| 명령 | 정지, 서, 위로 가, 아래로 가, 뒤로 가, 왼쪽으로 가, 오른쪽으로 가, 앞으로 가, 돌아, 하지 마, 그만해, 안 돼, 일어서, 앉아, 맘대로 해, 일어서, 앉아, 가지고 와, 가지고 가, 이리 와, 저쪽으로 가, 이쪽으로 가 |
| 감정 | 이쁜 짓 해봐, 윙크해 봐, 춤춰 봐, 착하지, 괜찮니?, 좋아, 잘 했어, 못 했어, 사랑해, 예쁘다 혼날래?, 어디 아프니?, 한번 더, 조용히 해 |
| 생활정보 | 오늘 날씨를 알려줘, 비오니?, 시원하니?, 온도가 몇 도야? |
| 날짜/시간 | 오늘은 몇 월 며칠이야?, 지금 시간이 몇 시지? |

## 3.2. 감정인식 실험 결과

이 논문은 지능로봇과 인간 사이의 음성인터페이스는 언어의 정보만을 전달하는 것이 아니라 음성에 포함되어 있는 감정을 인식하여 현재의 발화한 사람의 상태에 따라 대응하는 것으로 전제하고 있다. "사용자등록"이라는 감정이 없는 상태로 사용자를 등록한 후, 사용하는 것을 시나리오로 한다. 사람마다 발화시 운율 요소들이 변화하므로 기준이 될 수 있는 요소-감정이 없는 상태의 단어를 입력한다. 감정 음성 데이터베이스를 구성하는 단어나 문장에 감정이 잘 표현되었는지를 알아보기 위해 30명의 화자 중에서 29명(8700개)을 훈련, 1명(300개)은 테스트하는 교차검증방법으로 SVM을 사용하여 판단한 결과 약 59%의 정확도를 알 수 있었다.

재구성한 감정 음성 데이터베이스는 데이터베이스에서 사용한 발화 중 길이가 짧은 단어 25개를 제외한 발화와 계산한 감정인식 정확도를 기준으로 가장 높은 인식률 화자 2명, 중간 정도의 인식률 화자 2명, 가장 낮은 인식률 화자 1명 남녀 5명씩 선택하였다. 재구성한 감정 음성 데이터베이스 남녀 각 5명, 화자 당 발화수는 25개이며 "사용자 등록" 항목을 포함하여 총 1520개이다. 재구성한 감정 음성 데이터베이스에서 사용한 발화들은 <표 2>와 같다.

실험은 사람의 판단에 의한 것과 SVM을 사용하여 감정을 분류하는 방법, 두 가지를 비교 분석하였다. 사람의 판단에 의한 경우는 6가지의 감정이 포함되어 문장을 들려주고 판단하였으며 판단하는 감정의 기준은 감정 음성 데이터베이스 항목 중 감정이 없는 상태의 제시 문장인 '사용자등록'과 '내 이름은 OOO입니다'를 들려주는 방법을 사용하였다. SVM을 사용한 경우는 10명의 화자 중에서 9명(1,350개)을 훈련, 1명(150개)은 테스트로 사용하는 교차검증방법을 사용하였다.

<표 2> 재구성한 감정 음성 데이터베이스 사용 단어

| 항목 | 단어 및 문장 |
|---|---|
| 사용자등록 | 사용자등록, 내 이름은 ○○○입니다. |
| 인사 | 잘 지냈어?, 보고 싶었어, 오랜만이야!, 뭐하고 놀았어?, 갔다 올게, 나중에 봐 |
| 명령 | 아래로 가, 왼쪽으로 가, 오른쪽으로 가, 앞으로 가, 그만해, 맘대로 해, 가지고 와, 가지고 가, 저쪽으로 가, 이쪽으로 가 |
| 감정 | 이쁜 짓 해봐, 윙크해 봐, 어디 아프니?, 조용히 해 |
| 생활정보 | 오늘 날씨를 알려줘, 시원하니?, 온도가 몇 도야? |
| 날짜/시간 | 오늘은 몇 월 며칠이야?, 지금 시간이 몇 시지? |

<표 3>은 재구성한 감정 음성 데이터베이스를 구성하는 발화 1500개를 10명의 사람이 판단한 결과 값을 가지고 재정의한 감정인식결과를 입력으로 한 SVM에 의한 confusion matrix이다. 인식결과는 교차 검증으로 약 48%의 결과를 보여 감정을 분류하는데 있어 효과적임을 알 수 있었고, 사람에 의해서 판단한 결과는 약 71%의 정확도를 보였다. 감정을 분류하는 기준에서 SVM에 의한 감정인식기는 테스트 데이터에 대해 일정한 기준을 가지고 감정 상태를 분류하는 반면, 사람의 경우는 현재 몸 상태 또는 주변 상황에 따라 같은 감정 상태의 단어 또는 문장을 듣더라도 항상 같은 감정으로 분류할 수 없다. 그러나 사람의 의한 감정 분류는 화자가 음성에 표출한 다양한 감정을 복합적으로 한 개 이상의 감정이 있더라도 한 개 이상의 감정을 인식할 수 있다.

<표 3> SVM에 의한 감정인식 confusion matrix (%)

(평균 정확도: 47.6%)

| | angry | bored | happy | neutral | sad | surprised |
|---|---|---|---|---|---|---|
| angry | 48.5 | 2.4 | 7.6 | 17.1 | 1.6 | 22.8 |
| bored | 1.2 | 71.4 | 1.9 | 12.1 | 13.5 | 0.0 |
| happy | 12.2 | 7.8 | 39.0 | 26.6 | 2.0 | 12.4 |
| neutral | 9.1 | 12.1 | 17.5 | 50.1 | 5.5 | 5.6 |
| sad | 0.4 | 61.9 | 5.2 | 11.6 | 20.5 | 0.4 |
| surprised | 27.5 | 1.2 | 13.1 | 4.9 | 0.08 | 52.4 |

<표 4> 사람 판단에 의한 감정 분류의 confusion matrix (%)
(평균 정확도: 70.6%)

| | angry | bored | happy | neutral | sad | surprised |
|---|---|---|---|---|---|---|
| angry | 77.6 | 1.4 | 2.8 | 8.4 | 0.9 | 8.4 |
| bored | 1.4 | 55.4 | 2.8 | 5.5 | 35.4 | 0.2 |
| happy | 4.0 | 2.9 | 77.7 | 5.8 | 2.8 | 6.3 |
| neutral | 8.3 | 2.4 | 8.9 | 74.8 | 4.6 | 2.1 |
| sad | 1.0 | 37.4 | 2.3 | 2.9 | 55.5 | 0.4 |
| surprised | 7.8 | 0.6 | 5.4 | 2.6 | 0.8 | 82.6 |

<표 4>에서 사람 판단에 의한 감정 분류는 지루함과 슬픔의 경우 분류하는데 있어 오판이 생기는데 이 경우 피치와 에너지가 다른 감정보다 낮게 나타나는 경향이 있음을 알 수 있었다. 사람의 판단에 의한 감정 분류에서 청자가 남자인 경우 약 67%의 인식률인데 반해 여자의 경우는 약 80%의 인식률을 보임을 알 수 있었다.

피치, 에너지, Teager 에너지, 포만트의 특징을 각각 한 가지만 사용하여 감정을 인식한 결과는 피치의 경우 45%, 에너지의 경우 40%, Teager 에너지의 경우 39% 그리고 포만트(F1, F2, F3)의 경우 29%였다. 전체적으로 인식률이 가장 높은 것은 피치였다. 단일 특징을 사용하여 분류한 결과, 피치는 화남, 지루함과 기쁨, 에너지의 경우는 화남과 지루함, Teager 에너지의 경우는 지루함과 놀람을 그리고 포만트를 사용한 경우는 놀람을 분류하는데 있어 유효한 특징들이었다.

## 4. 결 론

이 논문은 지능로봇과 음성 인터페이스시 음성에 포함되어 있는 감정을 인식하는 시뮬레이션 결과를 기술하였다. 입력된 음성의 음향 및 운율 정보만을 사용하여 감정을 분류하였다. 사람이 감정을 음성으로 표현하거나 화자가 아닌 다른 사람이 느끼는 정도는 매우 주관적이다. 감정을 표현하는 정도는 대체적으로 개인적인 편차가 있지만 일반적으로 사람들간에 공통되는 점은 화나거나 기쁜 경우에는 피치와 에너지가 높게 나타나고 지루한 경우는 피치와 에너지가 낮게 나타났다. 놀람의 경우에는 포만트 주파수가 유효한 특징임을 알 수 있었다.

단일 특징을 사용한 감정인식 실험에서 피치, 에너지, Teager 에너지가 약 40% 정도의 정확도로 포만트는 약 29%의 정확도를 보인 반면, 주파수변이, 진폭변이는 인식률에 대한 기여도가 크지 않았다. 인식률을 향상하기 위해서 특징들을 조합하

여 유효한 감정인식 특징들을 추출하거나 피치를 추출하는데 있어 음성의 발화 방법 예측 또는 발화율에 따라 가변적인 피치 추출 알고리즘을 적용하는 것이 필요하다.

## 참 고 문 헌

[1] Moriyama and S. Ozawa, "Emotion recognition and synthesis system on speech", *IEEE Internationl Conference on Multimedia Computing and Systems*, pp. 840-844, 1999.

[2] O.-W. Kwon, K. Chan et al., "Emotion recognition by speech singnals", *Proc. Eurospeech*, Geneva, Switzerland, pp. 125-128, 2003.

[3] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combing acoustic features and linguistic information in hybrid support vector machine-belief network architecture", *Proc. ICASSP*, Montreal, Canada, pp. 577-580, 2004.

[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov Mdoel-based speech emotion recognition", *Proc. ICASSP*, Hongkong, China, pp. 401-404, 2003.

[6] T.-L. Pao and Y.-T. Chen, "Mandarin emotion recognition in speech", *IEEE workshop on Automatic Speech Recognition and Understanding*, pp. 227-230, 1999.

[7] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", *Proc. ICASSP*, Albuquerque, NM, pp. 381-384, 1990.

[8] G. S. Ying, L .H. Jamieson, and C. D. Michell, "A probabilistic approach to AMDF pitch detection", *Proc. ICSLP*, Philadelphia, PA, pp. 1201-1204, 1996.

[9] R. E. Slyh, W. T. Nelson, and E. G. Hansen, "Analysis of mrate, shimmer, jitter, and F0 contour features across stress and speaking style in the SUSAS database", *Proc. ICASSP*, Phoenix, AZ, pp. 2091-2094, 1999.

[10] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[11] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.

접수일자: 2005년 11월 23일
게재결정: 2005년 12월 23일

▶ 장광동(Kwang-Dong Jang)
주소: 361-763 충청북도 청주시 흥덕구 개신동 12번지
소속: 충북대학교 제어계측공학과
전화: 043)261-3374
E-mail: kdjang@chungbuk.ac.kr

▶ 김남(Nam Kim)
주소: 361-763 충청북도 청주시 흥덕구 개신동 12번지
소속: 충북대학교 전기전자컴퓨터공학부
전화: 043)261-2482
E-mail: nkim@chungbuk.ac.kr

▶ 권오욱(Oh-Wook Kwon) : 교신저자
주소: 361-763 충청북도 청주시 흥덕구 개신동 12번지
소속: 충북대학교 전기전자컴퓨터공학부
전화: 043)261-3374
E-mail: owkwon@chungbuk.ac.kr