500단어급 핵심어 검출기에서 화자적응 성능 평가*

서현철(전남대), 이경록(전남대), 김진영(전남대), 최승호(동신대)

<차 례>

<Abstract>

Speaker Adaptation Performance Evaluation in Keyword Spotting System

Hyun-Chul Seo, Kyong-Rok Lee, Jin-Young Kim, Seung-Ho Choi

This study presents performance analysis results of speaker adaptation for keyword spotting system. In this paper, we implemented MLLR (Maximum Likelihood Linear Regression) method on our middle size vocabulary keyword spotting system. This system was developed for directory services of universities and colleges. The experimental results show that speaker adaptation reduces the false alarm rate to 1/3 with the preservation of the mis-detection ratio. This improvement is achieved when speaker adaptation is applied to not only keyword models but also non-keyword models.

* 주제어: 음성인식, 핵심어 검출, 토큰 패싱 알고리듬, 화자적응, MLLR

* 본 연구는 2002년도 전남대학교 RRC HECS 연구비에 의하여 지원을 받았음.

1. 서 론

인간의 의사소통 수단인 음성은 사용의 편이성 및 자연성 면에서 다른 인터페이스에 비해 우수하며, 이런 음성을 실시간으로 처리하고 그 뜻을 파악하는 음성인식에 관한 연구가 근래에 들어 크게 활성화되고 있다. 이러한 음성인식은포괄적으로 고립단어인식, 연속음인식, 핵심어 검출로 나눌 수 있다. 이 중 핵심어 검출은 발성된 전체 문장을 입력받아 미리 정의된 핵심어 사전과 비교하여 문장 중에 포함된 핵심어의 출현여부를 조사한다. 또한 음성인식을 인식 주체에 따라 구분해 보면 화자독립 시스템, 화자종속 시스템, 화자적응 시스템으로 분류된다. 화자적응 시스템은 새로운 화자가 등장하면 새로 수집된 데이터를 기초로 기존 인식 시스템의 모델을 이 화자의 특성에 맞게 적응시키는 것이다[1]. 이는 일반적으로 어떤 화자에게도 좋은 성능을 갖는 시스템을 기대하지만 경험적으로 그렇지 못하기 때문이다. 화자적응의 기본 개념은 가능한 최소의 적응데이터를 사용하여 화자 독립 파라미터를 가능한 한 많은 화자 특정 정보를 갖는 화자 종속 파라미터로 변환시키는 것이다[2].

본 논문에서는 기존의 화자적응 알고리듬을 핵심어 검출기에 적용하여 그 성능을 평가함을 목적으로 한다. 사용된 핵심어 검출기는 대학구내 전화안내 시스템을 위해 개발된 중규모급 핵심어 검출기(약 500여개의 핵심어)를 사용하였다. 현재, 개발된 핵심어검출기는 핵심어의 개수가 많아, 아주 우수한 성능을 보이고 있지는 않지만, 화자적응을 이용하여 성능을 개선할 것으로 기대하였다.

본 논문에서 화자적응 기법으로 사용한 방법은 maximum likelihood linear regression(MLLR)이다. MLLR은 maximum a posteriori 기반의 적응과 같이 모델 파라미터를 적응시키는 변환기반 방법이지만, 변환 공유를 이용하여 적은 양의 데이터로부터 충분히 강인한 효과를 얻는 장점을 가지고 있다. 실험분석은 적응데이터의 양과 MLLR 적용시 클래스의 변화에 따라 이루어졌으며, 적응 후 핵심어 검출기의 성능을 화자적응 전의 성능과 비교하였다.

2. 핵심어 검출기의 개요

본 논문에서는 대학구내 전화안내 시스템을 위한 중규모급 핵심어 검출기를 사용하였다. 핵심어 검출 방식은 현재 가장 많이 사용되고 있는 핵심어 모델과 필러(filler) 모델을 선형 결합한 인식 네트워크를 이용하였다. 핵심어 모델은 552개의 단어모델로 이루어져있고, 필러 모델은 46개의 음소 모델과 4개의 비음성 모델로 이루어져있다. 핵심어 모델과 필러 모델은 토큰 패싱(token passing) 알고리듬을 적

용한 선형 네크워크를 이용하여 결합된다.

2.1. 음성 데이터 베이스

훈련 데이터 베이스는 남성화자 30명을 대상으로 1301문장을 발성하여 데이터를 구축하였고, 평가용 데이터 베이스는 훈련용 화자에 속하지 않은 새로운 남성화자 3명을 대상으로 496문장을 발성한 데이터로 구성된다. 녹음환경은 조용한 사무실 환경이고, 8kHz, 16bit로 A/D변환하였다.

2.1.1. 훈련용 데이터 베이스

훈련 데이터 베이스는 교내 전화안내 시스템에 적합한 552개의 핵심어와 실제 전화번호를 문의하는 통화 과정에서 자주 사용되는 56개의 비핵심어를 이용하였다. 핵심어는 특성에 따라 3단계(상위, 중위, 하위)로 구분하였다. 구분된 각 단계의 정보는 대학명, 대학의 기관명, 대학의 학과명과 부서명으로 구성된다. 비핵심어는 문장에서의 위치에 따라 3가지(문두, 문중, 문미)로 구분하였다.

훈련용 문장 선정에 있어서는 핵심어와 비핵심어의 출현 가능한 모든 형태를 고려하여 문장 코퍼스를 생성하였다. 생성된 문장 코퍼스를 트라이폰 정보를 바탕으로 정렬한 후, 모든 트라이폰을 포함하는 1,301문장을 선정하였다. 녹음 문장의 기본 문법구조는 아래와 같다.

HS(문두)+FDN(상위)+BS(문중)+SDN(중위)+BS(문중)+TDN(하위)+ES(문미)
여보세요+전남대학교+에 있는+공과대학+의+전자공학과+가 몇 번입니까?

<표 1> 텍스트 코퍼스 구축에 사용된 핵심어와 비핵심어 정보

구분	핵심어				비핵심어			
	상위	중위	하위	소계	문두	문중	문미	소계
수	366	20	166	552	18	3	35	56

2.1.2. 평가용 데이터 베이스

평가용 데이터 베이스는 문장 당 최대 3개까지의 핵심어를 포함하는 496문장을 선정하였다. 평가용 문장 중 핵심어가 포함되지 않는 문장은 비핵심어로 증권상장사 정보를 사용하였다. 핵심어를 포함하는 문장에서 핵심어의 출현은 최소 1회 이상 출현하도록 하였고, 출현한 핵심어의 총 수는 828개이다.

2.2. 핵심어 검출 시스템

2.2.1. 특징 파라미터

핵심어 검출기에서 사용한 특징 파라미터는 귀의 비선형적인 특성을 고려해 Mel-scale로 warping시킨 12차의 Mel-scale cepstrum과 1차의 Normalized log energy, 12차 delta-cepsturm과 1차 delta energy를 특징파라미터로 사용하였다. 또한 26차 리프터링을 통해 가중치를 부여하여 핵심어 검출기에 사용되는 HMM 파라미터의 초기화 입력으로 사용하였다.

2.2.2. 핵심어 검출 시스템 구조

핵심어 검출이란 연속적인 화자의 입력을 분석하여 사전에 정의된 핵심어를 검출하는 것이다. 검출부는 핵심어와 필러 모델을 선형 결합한 인식 네트워크를 이용하였다[3]. 핵심어 모델은 문맥 종속형 트라이폰 모델로 훈련DB 중 핵심어 출현 구간에 대해서 HTK(Hidden markov Tool Kit)를 이용하여 훈련되었고, 필러 모델은 문맥 독립형 음소 모델로 46개의 음소 모델과 4개의 비음성 모델을 사용하였다. 필러모델 또한 HTK를 이용하여 훈련DB 중 비핵심어 구간에 대해서 훈련을 하였다. 핵심어 모델들과 필러 모델들은 토큰 패싱 알고리듬을 적용한 선형 네트워크를 이용하여 결합된다[4].

핵심어 모델 네트워크는 사전에 정의된 상대적으로 중요한 의미를 가진 핵심어를 검출하는 역할을 한다. 이러한 핵심어 모델 네트워크는 각 핵심어 모델을 토큰 패싱 알고리듬을 이용한 선형 결합 네트워크로 구성된다. 핵심어의 인식은 핵심어 사전 상의 단어모델의 패턴과 입력음성의 유사도를 이용하였다.

필러 모델 네트워크는 연속된 화자의 발성 중 상대적으로 정보가 적은 비핵심어 구간을 처리한다. 필러 모델 네트워크는 단음소 모델들과 비음성부 모델들의 선형 결합으로 이루어져 있다. 핵심어 모델 네트워크와 동일하게 필러 모델 네트워크도 토큰 패싱 알고리듬을 적용하였다.

3. MLLR을 이용한 화자적응

본 논문에서는, 여러 가지 다양한 적응방법 중에서 스펙트랄 매핑(spectral mapping) 방법과 모델 매핑 방법의 단점을 보완한 MLLR을 이용하여 화자적응을 실행하였다[5].

3.1. 화자적응 개요

지난 수년 동안 화자독립 인식 시스템의 성능에 많은 진전이 있었다. 그러나 성능이 좋은 화자독립 시스템이라도 어떤 화자들에 있어서는 좋은 결과를 얻지 못한다. 일반적으로 특정 화자의 훈련데이터가 충분한 화자종속 시스템이 화자독립 시스템의 인식결과보다 더 좋다는 것으로 알려져 있다. 하지만 이러한 화자종속 시스템은 많은 데이터 양과 등록 시간의 지연으로 여러 분야에 있어서 바람직하지 않다. 따라서 기존의 음성 인식 시스템을 새로운 화자에 따라 조절하는 화자적응 기술이 큰 관심사이다.

3.2. Maximum Likelihood Linear Regression

3.2.1. MLLR 개요

MLLR은 적응데이터를 이용하여 HMM 파라미터를 조절하는데 사용되는 회귀기반 변환 행렬을 생성한다. 또한, MLLR은 변환 행렬의 공유를 이용하여 적응데이터에 나타나지 않는 모델들에 대해서도 강인한 적응 변환 행렬을 만들 수 있다. 이것은 제한된 적응데이터의 문제를 해결하는데 도움을 준다. 일반적으로, 음향공간상에서 화자들간의 주요 차이점은 음소들의 평균적인 위치에 있다고 생각하므로 MLLR에서는 주로 파라미터의 평균값을 적응시키는 평균적응이 수행된다[6].

3.2.2. Transform Sharing

화자독립 모델과 화자종속 모델의 사이의 모든 차이점을 정확히 찾기 위해서는 HMM 시스템의 개개의 가우시안 모델마다 하나의 적응 변환 행렬을 사용하면 된다. 하지만, 이렇게 적응된 모델을 정확히 추정하기 위해서는 너무 많은 양의 적응 데이터가 요구된다. 이러한 이유로 가우시안 모델들을 하나로 묶어서 한 개의 변환 행렬로 그 집합을 적응시키는 변환 행렬 공유를 이용한다[7]. 적응 데이터에 출현하지 않거나 데이터가 부족한 모델들에 대해서 그와 유사한 모델들의 변환

행렬을 적용하여 적응이 이루어진다.

3.2.3. Mean Transform Estimation

n차의 벡터인 single mixture 성분의 평균을 μ_s 로 정의하면, 적응된 평균값은 아래와 같이 추정된다.

(1) $$\hat{\mu}_s \quad W_s \xi_s$$

W_s는 mixture 성분 s 에 대한 $n \times (n+1)$ 변환 행렬이고, $\hat{\mu}_s$는 적응된 모델의 평균 벡터이며, $\xi_s = [w, \mu_{s_1}, \ldots, \mu_{s_n}]^t$ 는 확장된 평균 벡터이다. w는 offset term으로 $w=1$이면 offset, $w=0$이면 no offset 이다.

변환후 n차의 관측벡터 o를 생성하는 상태 s의 확률 밀도 함수는 다음과 같다.

(2)

$$b_s(o) = \frac{1}{(2\pi)^{\frac{n}{2}} |\sum_s|^{\frac{1}{2}}} \quad e^{-\frac{1}{2}(o - W_s \xi_s)' \sum_s^{-1} (o - W_s \xi_s)}$$

관측된 특징 벡터는 $O = o_1 \ldots o_T$ 와 같이 T frame으로 구성되며, 특징 벡터 O를 관측하는 동안 시간 t에서 상태 s가 점유할 확률은 아래와 같다.

(3) $$\gamma_s(t) = \frac{1}{P(O|\lambda)} \sum_{\theta \in \Theta} P(O, \theta_t = s|\lambda)$$

이때, $P(O, \theta_t = s|\lambda)$은 O를 생성하며 시간 t에서 상태 s가 점유하는 likelihood이고, $P(O|\lambda)$는 관측열을 생성하는 모델의 total likelihood이다. $\gamma_s(t)$는 forward-backward 알고리듬을 이용하여 계산된다.

모든 가우시안의 covariance matrix가 diagonal하다고 가정하고 R개의 가우시안 $s_1 \ldots s_R$이 W_s를 공유한다고 가정하면 최종적으로 구해야 할 변환 행렬 W_s는 아래의 식으로 표현된다.

(4) $$w_i = G_i^{-1} z_i$$

식(5)에서 z_i는 아래와 같은 행렬 Z 의 i 번째 열이다.

$$Z = \sum_{t=1}^{T} \sum_{r=1}^{R} \gamma_{s,r} \sum_{s,r}^{-1} o_t \xi'_{s,r} \tag{5}$$

그리고 G_i는

$$G_i = \sum_{r=1}^{R} c_{ii}^{(r)} \xi_{s,r} \xi'_{s,r} \tag{6}$$

이런 결과적인 식들의 전체 내용은 [8]에서 주어진 것들이다.

3.3. 실험 결과

3.3.1. 화자적응 전의 핵심어 검출 실험

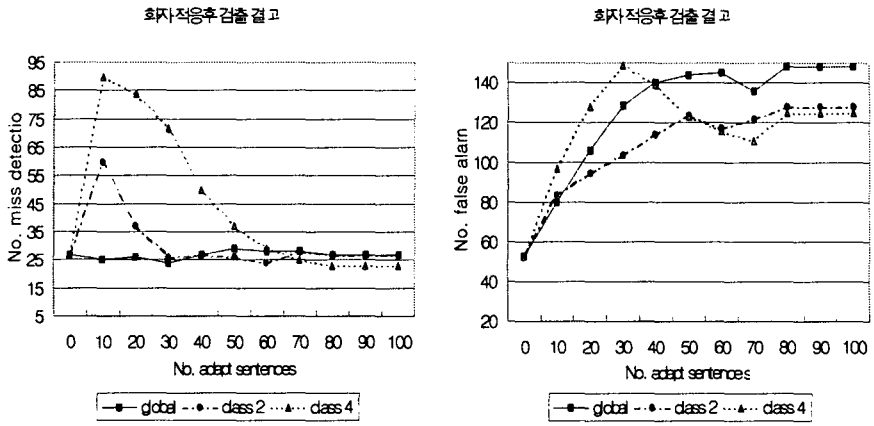
화자적응 적용전의 핵심어 검출 실험은 테스트 데이터 중 100문장을 핵심어 검출기에 참여하지 않은 화자로부터 녹음하여 사용했다. 핵심어 모델은 mixture 3, 필러 모델은 mixture 6을 사용하여 실험하였다. 검출결과는 핵심어를 인식하지 못하거나 다른 핵심어로 인식한 MDR(Missed Detection Rate), 핵심어가 아닌 구간에서 핵심어가 출현하거나 오인식한 핵심어를 FAR(False Alarm Rate)로 나타내었다. MDR과 FAR은 낮을수록 좋은 결과를 의미한다.

훈련 데이터를 핵심어와 비핵심어로 분류하여 훈련시킨 핵심어 모델과 필러 모델을 사용하여 실험한 결과 MDR은 27%이며, FAR은 53%이다.

3.3.2. 화자적응 후의 핵심어 검출 실험

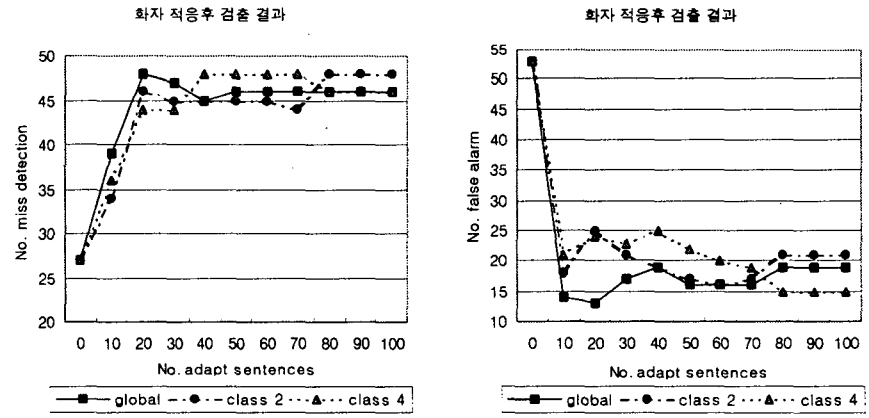
MLLR을 이용한 화자적응에 사용된 적응 데이터는 테스트 데이터 496문장중 임의의 100문장을 선정하였고 Supervised, batch mode로 기존 모델의 평균값만을 대상으로 적응하였다. 적응 대상으로는 핵심어 모델, 필러 모델, 핵심어와 필러모델 세 개로 구분하여 각 경우에 따라 적응데이터를 10에서 100까지 10문장씩 증가하며 변환 행렬을 생성하였다. 그리고 클래스 수의 변화에 따라 화자적응 성능을 평가하였다.

화자적응 후 핵심어 검출 실험에 사용한 데이터는 기존의 테스트 데이터 중에서 화자적응에 사용된 것을 제외한 나머지를 이용하였다.



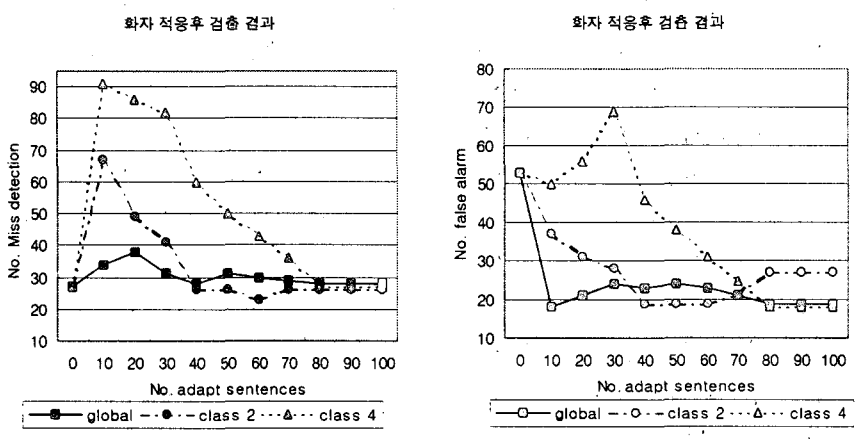
<그림 1> 핵심어 모델 적응 후 핵심어 검출 실험

핵심어 모델에만 화자적응을 적용한 그림 1의 결과는 적응 데이터의 증가에 따라 MDR의 성능이 화자적응 전과 비슷한 것을 보인다. 하지만 FAR의 경우에는 핵심어가 아닌 구간을 핵심어 모델로 인식되어 검출 결과가 좋지 않다.



<그림 2> 필러 모델 적응 후 핵심어 검출 실험

필러 모델만 적응시킨 그림 2의 결과를 보면 그림 1의 핵심어 모델을 적응한 실험과 상반되는 결과임을 알 수 있다. MDR의 경우는 결과가 아주 좋지 않지만 FAR은 상당히 줄어들었다. 필러 모델이 핵심어 모델을 잠식하여 핵심어 구간에서 검출되어야할 핵심어가 출현하지 않아 Missed Detection의 수가 두 배 가량 증가한 것으로 분석된다.



<그림 3> 핵심어 모델과 필러 모델 적응 후 핵심어 검출 실험

핵심어와 필러 모델을 개별적으로 적응했을 때 MDR과 FAR 어느 한쪽의 성능이 매우 좋지 않았다. 그래서, 핵심어 모델과 필러 모델을 모두 적응시켜 보면 적응 문장의 수가 증가함에 따라 MDR은 적응 전의 성능보다 향상되거나 비슷한 결과를 보였다. FAR은 클래스에 따라 조금은 다르지만 35% 정도 향상된 결과를 얻었다.

실험결과에서 보듯이 핵심어와 필러 모델을 모두 적응했을 때 적응데이터가 증가할수록 MDR의 성능은 적응전의 성능을 유지하며 FAR은 1/3로 줄어 성능이 향상됨을 알 수 있다. 적응데이터가 증가함에 따라 수렴하는 핵심어 검출 인식률은 아래와 같다.

<표 2> 각 모델별 화자적응 후 수렴되는 핵심어 검출 결과

적응대상	핵심어, 비핵심어			핵심어			비핵심어			비고
클래스 수	1	2	4	1	2	4	1	2	4	(적응전)
MDR(%)	28	26	27	27	27	23	46	48	46	27
FAR(%)	19	27	18	148	128	125	19	21	15	53

4. 결 론

핵심어 검출기의 성능 향상을 위해 본 논문에서는 화자적응 방법을 적용하였다. 많은 화자적응 방법 중 널리 쓰이고 있는 MLLR을 적용하여 기존 핵심어 검출기

의 성능과 화자적응 후의 성능을 평가하였다. 실험 결과에서 보았듯이 MDR은 뚜렷한 성능 향상이 없었다. 그러나 핵심어와 필러 모델을 모두 적응한 실험에서 보면 적응 데이터의 증가에 따라 MDR은 27%로 적응 전의 성능을 유지하며, FAR은 적응 전의 53%에서 35% 감소한 18%로 매우 좋은 결과를 얻을 수 있었다. 또한 클래스의 수가 증가할수록 요구되는 적응 데이터의 양이 증가하는 것을 볼 수 있었다. 이러한 이유는 클래스의 수가 증가할수록 데이터의 부족현상이 커지기 때문이다. 본 연구의 결과는 기본 핵심어 검출기의 성능이 우수하지는 않기 때문에, 연구결과의 정량적 수치가 큰 의미를 가진다고는 할 수 없겠으나, 화자적응이 핵심어 검출기에서 핵심어 및 필러 모델 모두에게 적용되어야 한다는 점을 밝혔으며, MDR의 개선보다는 FAR의 개선에 큰 효과가 있음을 밝혀냈다는 점에서 의미를 가진다고 하겠다.

앞으로 핵심어검출기의 성능개선, 새로운 클래스 분류 방법 및 핵심어 검출기에 적합한 화자적응 알고리듬을 연구해나갈 계획이다.

참 고 문 헌

[1] Leggetter, C. J. and P. C. Woodland (1995), Flexible Speaker Adaptation for large vocabulary speech recognition, In *Proceedings Eurospeech*, pp.1155~1158.

[2] Hao, Y. and D. Fang (1994), Speech Recognition Using Speaker Adaptation by System Parameter Transformation, *IEEE Transactions on Speech and Audio Processing* Vol.2 No.1 Part 1, January, pp.63~68.

[3] Junkawitsch, Jochen, Gunther Ruske and Harald Hoge (1997), Efficient methods for detecting keywords in continuous speech, *Proc. Eurospeech* 97. Vol.1, pp.259~262.

[4] Young, S. J., N. H. Russell and J. H. S. Thornton (1989), *Token Passing: a simple conceptual model for connected speech recognition systems*, Cambridge University Engineering Department.

[5] Hamaker, J. E. (1999), *MLLR: A Speaker Adaptation technique for LVCSR*, Mississippi State University Electrical and Computer Engineering Department.

[6] Leggetter, C. J. and P. C. Woodland (1995), Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models, *Computer Speech and Language* Vol.9, pp.171~185.

[7] Gales, M. J. F. (1996), The generation and use of regression class trees for MLLR adaptation, Technical Report, Cambridge University Engineering Department, August.

[8] Christensen, Heidi (1996), *Speaker Adaptation of Hidden Markov Models using Maximum Likelihood Linear Regression*, Aalborg University Institute of Electronic Systems Department of Communication Technology.

접수일자: 2002년 5월 3일
게재결정: 2002년 5월 24일

▶ 서현철(Hyun-Chul Seo)
주소: 500-757 광주광역시 북구 용봉동 300 전남대학교 공과대학 전자공학과
소속: 전남대학교 공과대학 전자공학과 신호처리실험실
전화: 062) 530-0472
Fax: 062) 530-0472
E-mail: turck182@dsp.chonnam.ac.kr

▶ 이경록(Kyong-Rok Lee)
주소: 500-757 광주광역시 북구 용봉동 300 전남대학교 공과대학 전자공학과
소속: 전남대학교 공과대학 전자공학과 신호처리실험실
전화: 062) 530-0472
Fax: 062) 530-0472
E-mail: krlee@dsp.chonnam.ac.kr

▶ 김진영(Jin-Young Kim)
주소: 500-757 광주광역시 북구 용봉동 300 전남대학교 공과대학 전자공학과
소속: 전남대학교 공과대학 전자공학과 신호처리실험실
전화: 062) 530-1757
Fax: 062) 530-0472
E-mail: kimjin@dsp.chonnam.ac.kr

▶ 최승호(Seung-Ho Choi)
주소: 520-714 전남 나주시 대호동 252번지 동신대학교 정보통신공학부
소속: 동신대학교 정보과학대학 정보통신공학부
전화: 061) 330-3194
Fax: 061) 330-2209
E-mail: shchoi@white.dongshinu.ac.kr