

보안 시스템을 위한 음향 신호의 위치추적 및 혼합된 음성 신호의 화자 수 추정방법

김영섭*, 이광석**, 고시영***, 허강인*

요약

본 논문은 실시간 보안 시스템을 위한 음향 신호의 위치추적 및 다수의 화자에 의해 동시 발생된 혼합된 음성 신호에서의 화자 수 추정에 관한 연구이다. 음향 신호의 위치 추적은 두개의 저가 사운드카드 및 네개의 콘덴서 마이크를 이용하여 비용적인 측면을 고려하였으며, 시스템의 기하학적 구조와 NCC(Normalized Cross Correlation)을 이용하여 마이크론 간 의 소리 도달 시간차를 구하는 계산량을 줄였다. 뿐만 아니라 혼합된 음성 신호에서의 화자 수를 추정하기 위하여 순수한 음성신호의 세기 포락선이 4~5Hz주위에 판별 가능한 두드러진 모듈레이션 패턴을 가지고 있다는 성질을 이용하여 화자 수 확인을 위한 음성 특징을 추출하고, 가우시안 혼합 모델(GMM)을 이용하여 화자 수를 추정하였다.

Acoustic Signal Localization and Estimation Method for the Number of Speakers in Mixed Speech Signals for Security System

Young-Sub Kim*, Kwang-Seok Lee**, Si-Young Koh***, Kang-In Hur*

ABSTRACT

This paper is a study on acoustic signal localization and estimation method of speaker number in mixed speech signals radiated to many speakers for security system. A proposed system of acoustic signal localization has reduced a complexity of calculation on sound arrival time-gap between microphones using geometry structure and NCC(Normalized Cross Correlation). Additionally, we allow for economic terms using two low price sound card and four condenser microphones. And then, sound features is abstracted for confirmation of speaker number using a property (that) it have notable modulation pattern that intensity envelope of purest sound signal can be determined around the 4~5Hz for speaker number estimation in mixed sound signal and we estimate speaker numbers using GMM.

Key Words : Acoustic Signal Localization, Time Differences of Arrival, Modulation Spectrum, Estimating the Number of Speakers, GMM(Gaussian Mixture Model)

* 동아대학교 전자공학과 (✉yskim770202@gmail.com)

** 국립경남과학기술대학교 전자공학과

*** 경일대학교 전자공학과

· 제1저자(First Author) : 김영섭 · 교신저자(Correspondent Author) : 허강인

· 접수일(2011년 11월 10일), 수정일(1차 : 2011년 12월 9일), 게재확정일(2011년 12월 12일)

I. 서 론

현대사회는 국가 및 개인의 재산과 안전에 대한 관심이 높아짐에 따라 보안의 중용성과 범위가 점차 넓어져 가고 있다. 기존에 널리 사용되는 영상 감시 시스템은 카메라가 획득할 수 있는 시각적 범위가 한정되어 있어 광범위한 지역에서의 감시가 쉽지 않으며, 이를 보완하기 위해 다수의 카메라를 이용할 경우 시스템이 커진다는 단점을 가지고 있다. 그러나 음향 신호를 이용하여 보안 시스템을 구성 할 경우 다양한 환경 및 대상에 적용할 수 있으며, PTZ(Pan-Tilt-Zoom)카메라를 조합하여 시스템을 구성할 경우 시스템의 구성이 간단하다는 장점을 가지고 있다. 더욱이 마이크로폰으로 입력되는 신호가 음성 신호일 경우 혼합된 음성에서의 화자 수 검출을 통해 침입자의 수를 카운트 할 수 있으며, 이는 보안시스템 뿐만 아니라 음성인식 시스템에도 적용 가능하다. 본 논문에서 사용된 위치 추적시스템은 기존 시스템에 비해 비용적인 측면을 고려하여 두 개의 저가 사운드카드 및 네 개의 컨테이너 마이크로폰으로 구성하였으며, 시스템의 기하학적 구조를 반영하여 독립적인 각 센서에 두 개의 마이크로폰을 내장하여 최대 지연을 1msec로 한 Normalized Cross Correlation을 이용함으로써 마이크로폰 간의 소리 도달 시간차를 구하는 계산량을 줄였다. 혼합된 음성신호에서의 화자 수 추정은 순수한 음성신호의 Intensity Envelope이 4~5Hz주위에 독특한 모듈레이션 패턴을 가지고 있다는 성질을 이용하여 화자 수 확인을 위한 음성 특징을 추출하고, 가우시안 혼합 모델을 이용하여 화자 수를 추정하였다. 본 논문은 다음과 같이 구성되어 있다. 2장은 음향신호를 이용한 위치추적 방법에 대하여 설명하고, 3장은 혼합된 음성신호의 화자 수 추정, 4장은 시스템 개요와 실험을 통한 적용가능성을 확인한 뒤, 5장에서 결론을 맺는다.

II. 음향신호를 이용한 위치추적

음향 신호를 이용한 위치 추적을 위해서는 각 마이크로폰에 입력된 신호를 이용하여, 음원 발생 지점으로부터 각 마이크로폰까지에 대한 도착 시간차인 도달지연시간(TDOA : Time Differences of Arrival)을 추정해야하며, n 개의 마이크로폰을 이용할 경우 위치 추정 알고리즘에 따라 $n(n-1)/2$ 개의 도달지연 시간을 계산해야한다[1][2].

2.1 음향신호의 시간 지연검출

본 논문에서는 음원의 위치를 추정하기 위해 센서와 음원사이의 도달각만을 이용함으로써 두 개의 입력 센서에 대해 각 센서가 한 번의 도달지연시간만을 계산하도록 하였다. 또한 마이크로폰에 입력된 신호의 길이에 따른 $O(N^2)$ 의 계산량을 줄이기 위해, 입력센서 내의 두 마이크로폰 사이의 최대 지연이 1msec(f_s : 16KHz, d_{ij} : 25cm)를 넘지 않는다는 가정하에 정규화된 상호상관 함수(NCC: Normalized Cross Correlation) 식 (1)을 이용하였다.

$$R_{ij}(\tau) = \frac{\sum_k (i_k - m_i) \times (j_{k+\tau} - m_j)}{\sqrt{\sum_k (i_k - m_i)^2} \sqrt{\sum_k (j_k - m_j)^2}} \quad (1)$$

이 때, i 와 j 는 두 개의 마이크로폰에 수신된 신호, m_i 와 m_j 는 각 수신된 신호의 평균, τ 는 두 마이크로폰 간에 소리가 도달하는 상대적인 최대 지연 시간차를 나타낸다. 마지막으로 식 (1)을 이용하여 구한 상관계수에서 최대 상관을 구하면, 두 신호가 마이크로폰에 도달하는데 걸리는 샘플 차이를 구할 수 있다.

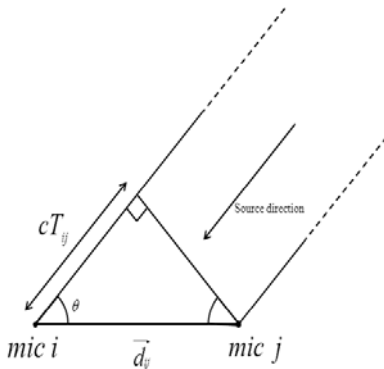


그림 1. 두 개의 마이크로폰과 음향신호의 관계
Fig 1. The relation between two microphones and acoustic signal

2.2 음향신호의 도달각 추정

도달지연 시간이 결정되면, cosine법칙을 이용하여 마이크로폰과 음향 신호의 발생 위치 사이의 도달각 (DOA : Direction of Arrival)을 계산할 수 있다. 그림 1은 입력센서 내의 두 마이크로폰과 음향 신호와의 관계를 기하학적으로 보여주며, θ 는 식 (2)과 같이 나타낼 수 있다.

$$\theta = \cos^{-1} \left(\frac{c \times n}{d_{ij} \times f_s} \right), c = 331.5 + 0.61t \quad (2)$$

이 때, c 는 음향신호의 속도, n 은 두 마이크로폰 사이의 샘플 차이, d_{ij} 는 두 마이크로폰 사이의 거리, f_s 는 샘플링 주파수이다. 그림 2는 각 입력센서의 도달각 추정을 위한 테스트 설정을 보여준다.

본 논문에서는 정확한 도달각 추정을 위하여 100Hz에서부터 5kHz까지의 주파수 범위를 가진 1초 길이의 chirp 신호와 실제 음성 신호를 5회 발생하여 그림 2와 같이 입력센서의 중심을 기준으로 90°와 45°에 대한 도달각을 검사하였다.

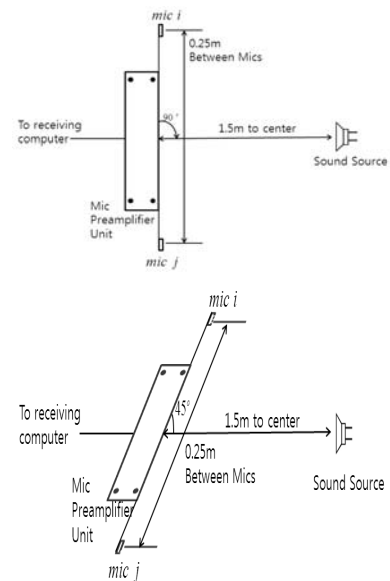


그림 2. 입력센서의 테스트 설정
Fig 2. Test set-up of input sensor

본 논문에서는 정확한 도달각 추정을 위하여 100Hz에서부터 5kHz까지의 주파수 범위를 가진 1초 길이의 chirp 신호와 실제 음성 신호를 5회 발생하여 그림 2와 같이 입력센서의 중심을 기준으로 90°와 45°에 대한 도달각을 검사하였다.

검사결과는 표 1과 같으며, 그림 3은 두 번째 테스트 구성으로 마이크로폰에 수신된 Chirp신호와 각 수신된 신호와 원신호의 상관신호, 그리고 상관 신호의 피크 쌍 중 하나를 확대하여 보여준다.

표 1. 테스트 설정에 대한 결과
Table 1. Results from each of our test configuration

	First Test Configuration			Second Test Configuration		
	Set-Up	Calculated	Sample Gap	Set-Up	Calculated	Sample Gap
Chirp Signal	90°	90°	0	45°	44.82°	23
Speech Signal	90°	90°	0	45°	44.82°	23

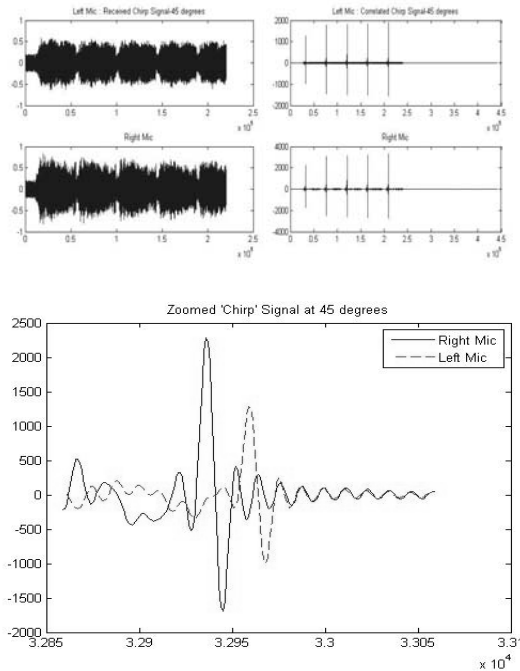


그림 3. 테스트 설정에 대한 출력신호
Fig 3. Output signals from our test configuration

마지막으로 추정된 도달각을 이용한 음향 신호의 위치 판별은 단순 감산 및 세 가지 조건(음향 신호가 두 개의 센서 좌측 또는 우측에 위치하거나, 두 개의 센서 사이에 위치할 경우)에 대해 이차원 평면에서의 정확한 좌표를 구할 수 있다.

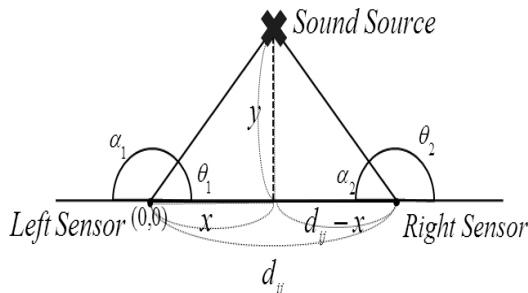


그림 4. 두 개의 입력센서와 음향신호의 관계
Fig 4. The Relation between two input-sensors and acoustic signal

그림 4와 식 (3)은 두 개의 센서 사이에 음향 신호가 위치할 경우에 대한 기하학적 관계 및 관련 수식을 보여준다.

$$x = \frac{d_{ij} \times \tan \alpha_2}{\tan \alpha_2 + \tan \theta_1}, y = x \cdot \tan \theta_1 \quad (3)$$

식 (4)와 식 (5)은 음향신호가 두 개 센서의 좌측과 우측에 위치할 때 x, y 좌표를 구하는 방법에 대해 설명한다.

$$x = \frac{d_{ij} \times \tan \alpha_2}{\tan \alpha_2 - \tan \alpha_1}, y = -x \cdot \tan \alpha_1 \quad (4)$$

$$x = \frac{d_{ij} \times \tan \theta_2}{\tan \theta_2 - \tan \theta_1}, y = x \cdot \tan \theta_1 \quad (5)$$

III. 혼합된 음성신호의 화자 수 추정

음성 신호는 종종 진폭 변조된 신호로서 모델링되며, 그 때 공진기(Resonator)와 음원(Sound Source)은 각각 변조기와 캐리어로서의 역할을 수행하며, 공진기의 주파수 특성은 시간의 함수로서 변화함으로 변조기 역시 시간에 동적인 특성을 갖게 된다. 또한 변조 주파수 4~5Hz성분이 일반적으로 우세하며, 그때 음성 특징들의 시간적 동적 특성(Temporal Dynamics)이 분석되어 진다[3][4].

반면에 사람의 청각 시스템은 음성의 시간적 동적 특성에 대해 대역 제한된 특징들을 가지고 있어서 1~16Hz사이의 변조 특징이 손실되면 음성의 명료성은 감소하게 된다는 것을 [5]를 통하여 알 수 있다.

3.1 모듈레이션 스펙트럼

화자 수에 대한 모듈레이션 스펙트럼은 음성 신호의 세기에 대한 포락선을 주파수 도메인에서 분석함으로써 계산되어지며, 8Hz아래에서 중요한 성분을 가지고 있다.

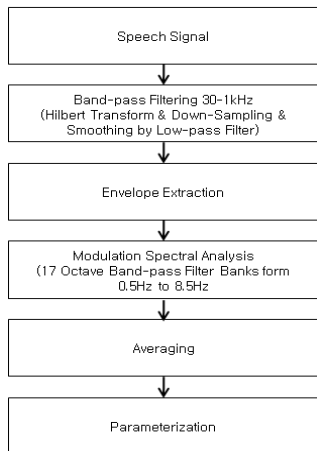


그림 5. 모듈레이션 특징 분석을 위한 블록 다이어그램
Fig 5. Block diagram for the modulation feature analysis

그림 5는 입력 신호에 대한 모듈레이션 특징을 분석하기 위한 과정이다. 입력 신호는 16kHz로 샘플링한 뒤 모듈레이션 특징의 안정성을 높이기 위해 30~1000Hz 대역통과 필터를 적용한다. 그리고 입력 신호의 강도에 대한 포락선을 구하기 위해서는 힐버트 변환(Hilbert Transform)[6]과 다운샘플링(Down Sampling)을 통하여 정확한 포락선 성분을 얻는다. 그리고 모듈레이션 스펙트럼 상에서 0.5Hz에서 8.5Hz사이의 중심주파수를 가지는 17차 옥타브 필터뱅크를 적용한 후, 그 출력신호들의 평균을 취하여 화자수의 분석용 특징으로 이용한다.

그림 6은 입력신호와 입력 신호에 대한 포락선, 그리고 모듈레이션 패턴을 나타낸다. 화자의 수가 증가함에 따라 그림 6에서와 같이 모듈레이션 특징 패턴은 3~5Hz주변에 피크가 한 사람이 발성한 음성신호에 대한 피크보다 작은 값을 나타냄을 확인할 수 있다. 이

러한 특징은 음성신호가 가지고 있는 위상의 비동기 성질에 기인한다. 그리고 화자의 수를 추정하기 위한 특징 파라미터는 스펙트럼 신호에 대한 누적 합, 피크의 최대값, 그리고 모듈레이션 패턴의 평균값을 분석함으로써 얻을 수 있는데, 화자의 수를 추정할 때 생기는 오류는 이들 특징들 간의 차이에 기인함을 알 수 있다.

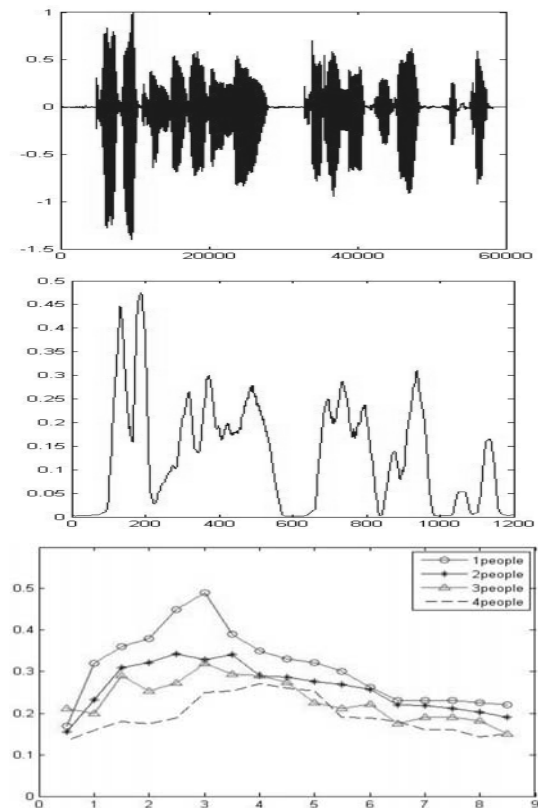


그림 6. 입력 신호의 모듈레이션 특징
Fig 6. Modulation feature of input signals

IV. 제안된 시스템 및 실험 결과

본 논문에서 구성한 실시간 위치 추적 및 화자 수 추정 시스템은 두 개의 마이크를 포함한 Left와 Right 센

서로 구성되며, 각 센서는 전치 증폭기를 거쳐 사운드 카드의 라인 입력을 통해 16kHz, 8bit의 분해능을 가진 음향 신호로 수집되었다.

그림 7과 그림 8은 실험에 대한 마이크 증폭모듈[7]과 시스템 구성을 보여준다. 음향신호의 위치추적은 그림 9와 같이 정확한 실험 결과를 위해 스피커를 통해 출력되는 음성신호를 Source Path위를 이동하면서 행해졌다. 결과는 10cm이내의 오차로 거의 정확한 값을 확인 할 수 있다.

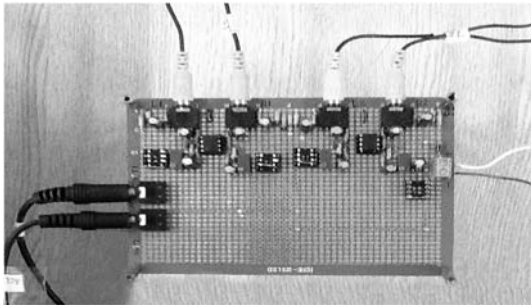


그림 7. 실험에 사용된 마이크 증폭 모듈[7]
Fig 7. Microphones amplifier module using experiment

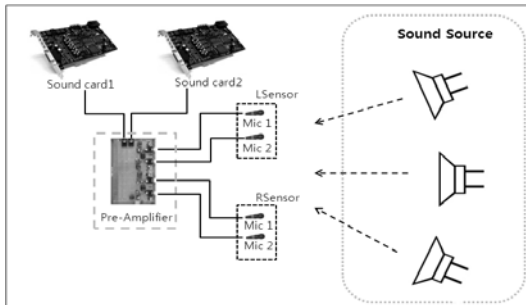


그림 8. 시스템 구성
Fig 8. System configuration

화자 수 추정에 대해서는 "ETRI 한국어 증가마이크 음성인식용 낭독체 문장" 데이터를 사용하여 1000번의 실험을 통한 분류 실험을 수행하였으며, 표 2는 화자 수에 대하여 모듈레이션 패턴으로부터 획득되어진

통계적 수치에 대한 특징 파라미터를 나타낸다.

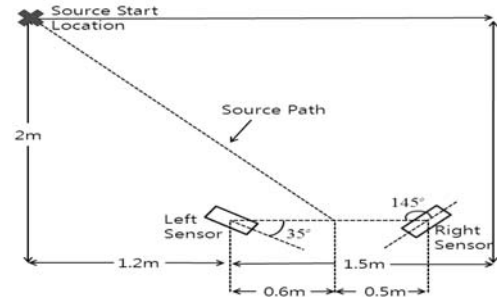


그림 9. 실험을 위한 마이크로폰 배치
Fig 9. Microphone configuration for experiment

표 3은 가우시안 혼합 모델(GMM)을 통하여 분석된 특징에 대한 분류 모델을 구축한 후, 혼합 음성에 대한 분류 실험 결과를 보여준다. 전체 실험에서 한 사람의 음성에 대한 분류는 70(%)이상의 결과를 보였으나, 다른 분류에서는 50(%)보다 적은 분류 결과를 확인할 수 있다.

표 3. 분류결과
Table 3. Classification Results

(%)	1	2	3	4
1	72.3	15.4	5.4	6.9
2	23.2	29.5	17.9	29.5
3	3.3	18.9	18.9	58.9
4	3.3	16.7	23.3	56.7

V. 결론

본 연구는 음향 신호의 위치추적 및 혼합된 음성 신호에서의 화자의 수를 검출하는 방법을 구현하였다. 음향 신호의 위치추적을 위해서는 두 개의 독립적인 입력 센서를 이용하여 각 센서가 한 번의 도달지연 시간만을 계산하도록 하였으며, 위치추정 또한 간단한

삼각합수를 이용하여 적은 계산량과 추정결과의 정확도를 확인하였다. 뿐만아니라 마이크의 수를 증가시킴으로서 3차원 음원위치 추적도 가능하다는 것을 알 수 있다. 또한 순수한 음성신호의 Intensity Envelope 이 4~5Hz주위에 독특한 모듈레이션 패턴을 가지고 있는 것에 기반한 화자수 추정은 분류 실험 결과 화자의 수 증가에 따라 분류 성능이 현저히 떨어지는 것을 확인할 수 있었지만, 입력되는 음성이 단일 음성인지 혼합 음성인지를 확인하는 데는 우수한 성능을 보임을 확인하였다.

참고문헌

- [1] C. Knapp and G.Carter, "The generalized correlation method for estimation of thime delay", Acoustics, Speech and Signal Processing, IEEE Transactions on, Vol. 24, no.4, pp.320-327, Aug. 1976.
- [2] 기창돈, 김강호, 이택진, "반향음과 잡음 환경을 고려한 실시간 소리 추적 시스템", 한국항공우주학회 한국항공우주학회지 제38권, 제3호, pp.258-263, 2010.
- [3] T.Houtgast and H.J.Steeneken, "A review of the MTF concept in room acoustics and its use of estimating speech intelligibility in auditoria", J.Acoust.Soc.Am., Vol.77, pp.34-41, 1995.
- [4] T.Arai and S.Greenberg, "The temporal properties of spoken Japanese are similar to those of English", Proc.of Eurospeech, Vol.2, pp.1011-1014, 1997.
- [5] R. Drullman, J.M.Festen and R.Plomp, "Effect of temporal envelope smearing of speech reception", J.Acoust.Soc.Am., Vol95, pp.1053-1064, 1994.
- [6] L.R.Rabiner, R.W.Schafer, Digital Processing of Speech Signals, PRENTICE HALL.
- [7] <http://circuitdiagram.net/stereo-electret-mic-preamplifier.html/>

감사의 글

본 논문은 동아대학교 학술연구비 지원에 의하여 연구되었음.

저자소개

김영섭(Young-Sub Kim)



2005년 동명정보대학교 컴퓨터공학과(공학사)
2007년 동아대학교 대학원 전자공학과(공학석사)

2009년~현재 동아대학교 대학원 전자공학과(박사과정)
※ 관심분야: 음성/영상신호처리, 패턴인식

이광석(Gwang-Seok Lee)



1983년 동아대학교 전자공학과(공학사)
1985년 동아대학교 대학원 전자공학과(공학석사)
1992년 동아대학교 대학원 전자공학과(공학박사)

1995년~현재 국립경남과학기술대학교 전자공학과 교수
※ 관심분야: 음성신호처리 및 인식, 퍼지 및 신경회로망, Biometrics, 지능화 기술

고시영(Si-Young Koh)



1979년 영남대학교 전자공학과(공학사)
1983년 영남대학교 대학원 전자공학과(공학석사)
1992년 동아대학교 대학원 전자공학과(공학박사)

1986년~현재 경일대학교 전자공학과 교수
※ 관심분야: 음성신호처리 및 인식, 지능화 기술

허강인(Kang-In Hur)



1998년~1989년 일본 쓰쿠바대학 객원 연구원
1992년~1993년 일본 도요하시대학 객원 연구원

1989년~현재 동아대학교 전자공학과 교수
※ 관심분야: 음성/영상신호처리, 패턴인식