# 순시적인 신호대 잡음비 예측과 RASTA 기법을 이용한 음성인식

# A Speech Recognition Using Instantaneous SNR Estimation and RASTA Processing

배 현 권*, 오 문 식*, 이 행 세**
(Hyun Kwon Bae*, Moon Sik Oh*, Haing Sei Lee**)

요 약

본 논문에서는 잡음에 강한 음성 인식기를 위한 음성의 특징 추출에 관해서 살펴 보았다. 지금까지의 음성 인식기는 조용한 실험실 환경하에서 학습이 이루어지나 실제 테스트는 여러 가지 환경에서 이루어지므로, 이러한 환경 변화에 따라 음성인식 시스템의 성능이 감소함을 보여왔다. 이를 보완하기 위해 여러 가지 연구가 진행되고 있으나 본 연구에서는 음성의 특징 추출 부분에서 순시적인 신호대 잡음비 예측과 잡음에 강한(RASTA)처리를 하므로써 인식율을 향상시켰다.

ABSTRACT

In this paper, we study a speech feature extraction method for noise-robust speech recognition. Generally recognizers have worked under ideal condition(noiseless), but we choose noisy environment in this experiment. We know this different environment typically degrades efficiency of recognizer, and many researchs have been performed to improve the effciency. We can reinforce recognition rate using RASTA processing and instantaneous SNR estimation which endures noise.

## Ⅰ. 서 론

지금 까지의 음성인식 기술은 실험실내에서 양질의 음성을 대상으로한 제한적인 것이 대부분이었다. 조용한 환경에서의 음성인식기는 이미 높은 성능을 보이고 있으나, 실제 환경에 존재하는 여러 요인에 의해 성능이 저하된다. 잡음에 의한 인식기의 성능저하는 인식기의 응용을 막는 요인의 하나로서 많은 연구가 진행되고 있다.[1][2] 잡음환경에 강한 음성인식을 위해 여러 가지 접근 방법이 연구되고 있다. 그 중 하나가 잡음의 첨가에 강인한 특징추출로서 RASTA (RelAtive SpecTrAl)처리가 있다. RASTA 처리 기술은 음성의 비음성적인 요소의 변화율이 성도 모양의 일반적인 변화율 밖에 놓여 있다는 사실에 기인한다. 따라서, 음성의 일반적인 변화율보다 더 느리거나 빠른 변화를 갖는 요소를 억제 하므로서 이루어진다.

본 논문에서는 RASTA 처리 기술과 잡음 음성신호의 즉각적인 신호대 잡음비를 예측하여 인지선형예측 모델과 조합한 새로운 RASTA PLP 분석법을 이용하였다. 음성 데이터 수집 및 실험 방법은 조용한 실험실 환경에서 데이터를 수집, 학습시키고 잡음 환경에서 수집한 음성과 잡음을 첨가한 음성을 가지고 인식 실험을 수행하여 기존의 음성 전처리 방법과 인식 성능을 비교하였다.

본 논문의 구성은 전부 6장으로 구성되어 있으며, 제 2장에서는 즉각적인 신호대 잡음비 예측 알고리듬과 RASTA 처리 기법에 대해서 기술하였다. 제 3장에서는 인식에 사용된 주 처리부와 후 처리부에 대해서, 제 4장에서는 실험환경과 데이터 구성에 대해서 설명하였다. 마지막으로 제 5장과 제 6장에서는 실험 결과 및 결론을 맺었다.

## Ⅱ. 음성의 특징 추출

본 연구에서 사용된 음성의 특징 추출은 인간의 청각구조를 공학적으로 모델링 한, PLP 분석법을 기초로 이루어 졌다. PLP 분석법은 음성의 고주파 해상도를 감소시키고 LPC 계수보다 적은 계수를 사용하여 더 좋은 인식율을 나타낸다.[3] 깨끗한 음성에서는 PLP 분석법이 필터링을 한 RASTA PLP 분석법보다 인식율이 좋다는 결과에 의해 신호대 잡음비에 따라 PLP 방법과 RASTA PLP 방법을 적절하게 적용하였다. 그림 1은 PLP를 이용한 본 논문에서 제안한 음성의 특징 추출과정을 나타낸 블록도이다.

*아주대학교 전자공학과
**아주대학교 전기전자공학부

접수일자 : 1997년 10월 22일

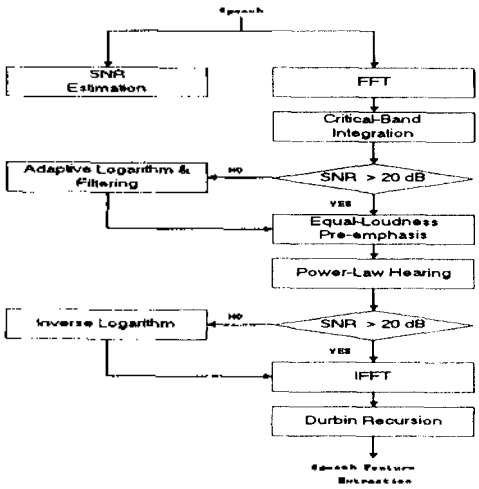



그림 1. 제안한 음성의 특징추출 과정

### 2.1 순시적인 신호대 잡음비 예측

일반적인 잡음 음성의 SNR예측은 음성활동 구간에서 단지 노이즈 하고만 관련된 부분을 이용하여 통계적으로 예측하여 왔다. 노이즈 파워 예측 값을 새로 수정하기 위한 모든 경우에는 음성 신호 세그먼트 구간에서 음성이 없는 구간이 필요하다. 그러나 사용한 알고리즘은 노이즈의 통계치를 수집하기 위해 음성과 비음성의 결정이 필요없고, 음성 활동 동안에 변하는 노이즈 단계를 쫓아갈수 있다.[4]

본 논문에서는 잡음 음성신호 $x(i)$는 대역 제한되고, 음성신호 $s(i)$와 잡음신호 $n(i)$의 합, $x(i)=s(i)+n(i)$이라고 가정했다. 여기에서 $i$는 시간을 나타낸다. 또한 $s(i)$와 $n(i)$는 통계적으로 독립이라고 가정하고, 따라서 $E\{x^2(i)\}=E\{s^2(i)\}+E\{n^2(i)\}$이다.

신호대 잡음비 $SNR_x(i)$의 계산은 주어진 윈도우 안의 단구간 파워 $\overline{p_x}(i)$의 최소값으로 얻을 수 있는 노이즈 파워 예측 $p_n(i)$에 기초를 둔다. $SNR_x(i)$는 시간 i에서 잡음 음성신호 $x(i)$의 신호대 잡음비 예측이다.

SNR예측은 다음의 세가지 단계로 수행한다.

1. 신호 x(i)의 부드러운 단구간 파워 예측 $\overline{P_x}(i)$의 계산.

$$P_x(i)=P_x(i-1)+x(i)*x(i)-x(i-N)*x(i-N)$$

$$\overline{P_x}(i)=\alpha*\overline{P_x}(i-1)+(1-\alpha)*P_x(i) \quad (1)$$

2. 노이즈 파워 예측 $P_n(i)$의 계산

$$P_n(i)=\min(\overline{P_x}(i),\ P_n(i)) \quad (2)$$

a. 천천히 변하는 노이즈 파워의 경우

$$P_n(i)=P_{Mmin}(i=r*M) \quad (3)$$

b. 빠르게 변하는 노이즈 파워의 경우

$$P_{Lmin}(i)=\min(P_{Mmin}(i=r*M),\ P_{Mmin}(i=(h-1)*\cdots,P_{Mmin}(i=(r-W+1)*M)) \quad (4)$$

3. $SNR_x(i)$의 계산

SNR의 예측은 예측된 최소 노이즈 파워 $P_n(i)$에 의해서 계산할 수 있다.

$$SNR(i)=10*\log_{10}\left(\frac{\overline{P_x}(i)-\min(\beta*P_n(i),\ \overline{P_x}(i))}{\beta*P_n(i)}\right) \quad (5)$$

여기서, $\beta$는 과예측 계수 (실제 실험에서는 주로 $\beta=1.5$ 사용)이다.

$SNR_x(i)$는 시간 $i$에서 음성 신호 $x(i)$의 신호대 잡음 비를 나타낸다. 그림 2는 음성 신호 "공일구칠"의 예측된 신호대 잡음비를 나타낸 히스토그램이다.

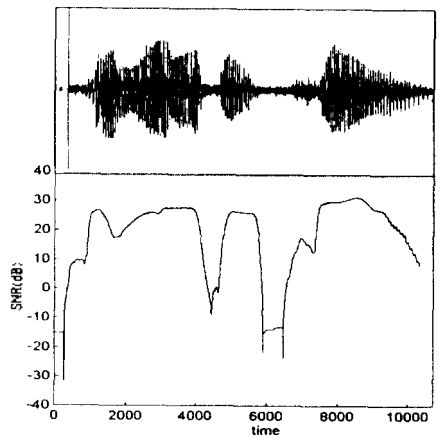



그림 2. 음성 파형 "공일구칠"과 SNR 예측

### 2.2 음성 파라미터의 대수 변환

음성 파라미터의 대수 변환은 FFT 변환 후에 시행하므로서 음성의 주 활동 영역이 아닌 고주파 부분을 신축한다.

음성 파라미터의 대수 변환과 역 대수 변환 수식

$$y=\ln(1+x/J) \quad (6)$$

여기서, $J$ 값은 전극 모델의 스펙트럼에 영향을 미치며 신호의 잡음 비에 따라 신축적으로 변한다. 역 대수 변환은 모든 y에 대해 양수를 보장 못하므로 근사적으로 식 (7)와 같이한다.

$$x=e^{(y)}/J \quad (7)$$

J 파라미터 계산:

a. 일반적인 방법

$$J = 1.0/(C \cdot E_n) \quad (8)$$

$C$ = 상수, $E_n$ = 평균 잡음 에너지

$E_n$ (평균 잡음 에너지)계산

처음에 들어오는 100ms의 음성 샘플들은 노이즈 하고만 관련이 있다고 가정(음성 데이터에서 처음 4프레임 부분은 실제 음성 데이터가 아니다.)하고 노이즈 power를 매 100ms마다 히스토그램 방법을 사용하여 계산한다.

b. SNR을 이용한 방법

예측된 SNR에 따라 J값을 변화시키면서 가장 좋은 인식율을 보이는 것을 해당 SNR의 J값으로 결정하였다. 표 5에 결과를 나타내었다.

### 2.3 RASTA 처리

RASTA 처리 기술은 음성의 비음성적인 요소의 변화율이 성도 모양의 일반적인 변화율 밖에 놓여있다는 사실에 기인한다. 따라서, 음성의 일반적인 변화율보다 더 느리거나 빠른 변화를 갖는 요소를 억제 하므로서 이루어진다. RASTA는 음성의 대수변환 파라미터들(대수변환 스펙트럼 에너지 혹은 켑스트럼 에너지)의 대역 통과 필터에 의해서 수행된다.

여과기의 변환 함수 (BPF)

$$H(z) = \frac{\alpha \sum_{n=0}^{N} (0 - \frac{N-1}{2}) z^{-n}}{1 - \rho z^{-1}} \quad (9)$$

$\alpha = 0.1$, $\rho = 0.9$, N = 5

그림 3은 대역 통과 필터의 주파수 특성이다.

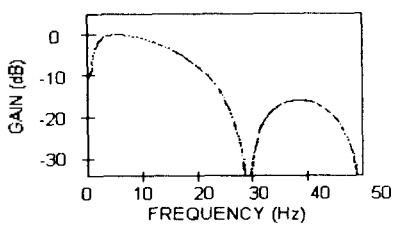
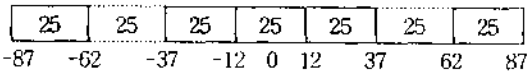



그림 3. BPF의 주파수 특성

## Ⅲ. 신경망과 마코프 모델

신경망은 음성에서 추출된 특징 계수를 이용하여 음소를 인식한다. 신경망 입력은 연속 음성에서 음소의 특성 변화를 얻기위해 현재 프레임을 중심으로 전후 프레임을 입력 벡터로 구성하였다. 프레임 인식은 역전파 알고즘을 이용하여 수행하였다. 한 프레임은 5개의 벡터로 구성되며, 벡터 하나는 5차의 PLP 혹은 제안한 RASTA PLP, 에너지, 영교차율로 구성되어 있다. 신경회로망의 학습과 인식을 위해 입력벡터를 다음과 같이 구성하였다.

단위 : ms




음소의 시간 변화를 관찰하기 위하여 175ms의 시간 구간에 대해 입력벡터를 구성하여 음소간의 관계 학습을 위해 구성하였다. Markov 모델은 통계적 특성 외에 시간적인 변화를 잘 반영시킬 수 있다. 음소열로 이루어진 단어는 자모의 순서와 밀접한 관련을 가지고 있다. 따라서 현재의 확률이 그 전에 발생한 데이타의 확률에 의존하는 Markov 모델은 가장 적정한 인식 방법이다.[5]

음성 신호 '공구'를 인식하였을 때 프레임 인식 결과는 ㄱㄱㄱㄱㄱㅗㅗㅗㅗㅗㅗㅗㅗㅗㅇㅇㅇㅇㅇㅇㅇ ㄱㄱㄱㄱ ㅜㅜㅜㅜㅜㅜㅜㅜ로 인식되어져 나오는데 이를 '공구'라는 문자로 바꾸어 준다. 이 때 자음은 갯수가 적고 모음은 많은 개수가 나타나는데 모음이 적은 수로 나타나면 잡음으로 보고 이를 제거하고 ㄴㄹㅁㅇ도 모음보다는 적지만 꽤 많은 갯수가 나타나므로 이것도 적게 나타나는 것은 잡음으로 보고 제거할 수 있다. 자음도 아주 짧게 나타나는 경우는 잡음으로 볼 수 있으므로 전구간에 걸쳐 독립적으로 한 개씩 나타나는 경우는 제거시킨다. 이렇게 바꾸어진 단어는 Markov 모델에 의해서 단어 인식 과정을 거치게 된다. '공구'라는 단어의 프레임 인식에 잡음이 섞여서 '공ㅅ구' 라는 문자로 인식되었을 때 이것이 '공수' 인지 아니면 '공구'인지의 구별을 하여야 하는데 이 때 통계적 처리과정을 거치는 Markov 모델이 확률적으로 구분해 주게 된다.

## Ⅳ. 인식 실험 및 데이터 구성

### 4.1 인식 시스템

음성 인식을 위하여 일반 실험실 환경에서 발음한 숫자음을 DT2801A 음성보드를 이용하여 10kHz, 12bits로 양자화한 데이터를 각각 5차의 PLP, 제안한 RASTA PLP 계수를 추출한 뒤, 이를 음성인식기의 입력으로 사용하여 SUN SPARC10 에서 실험을 하였다.

본 논문에서는 신경회로망과 Markov 모델을 결합한 음성인식기를 사용한다.[6] 신경회로망은 각 음성상의 음소를 인식하기 위하여 사용하고, 신경회로망의 출력인 음소열은 Markov 모델링하여 단어 인식시 사용한다. 그림 4는 실험에 사용된 인식시스템을 나타낸다.

### 4.2 실험 데이터 구성

실험 데이터는 1자리 숫자 10개와 4자리 숫자로 이루

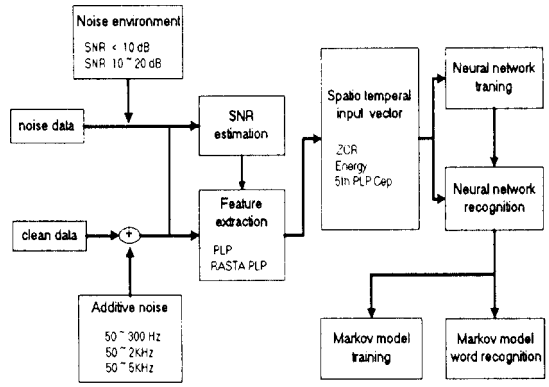



그림 4. 음성인식 시스템

어진 20개의 연속 숫자음을 사용한다. 신경회로망을 이용한 음소학습을 위하여 20대 남성 5인이 각각 2회씩 발음한 데이터를 사용하였고, 신경회로망을 통하여 인식된 음소열은 각 단어에 대한 Markov 모델을 구성하는데 사용되었다. 그리고 화자독립의 인식을 위해서 학습에 참여하지 않은 20대 남성 5인이 각각 1회씩 발음한 데이터, 총 10인이 발음한 데이터를 사용하여 인식 실험을 하였다. 잡음 데이터 수집은 청소기 잡음 및 각종 기기잡음을 테이프로 녹음하여, 환경 잡음으로 사용하고 실험 데이터를 구성하였다. 잡음 첨가 데이터는 주파수 대역이 다른 잡음 파형(그림 5에 보여준 선풍기, 진공 청소기, 전화 잡음)을 깨끗한 음성에 더해서 구성하였다. 또한 확장된 숫자음 인식을 위해 공부터 구구구구까지 10000개의 단어중에 50개의 단어를 임의로 추출해 인식실험을 하였다. 그림 6은 잡음 파형의 스펙트로그램을 나타낸다.

표 1. 실험에 사용한 데이터

| |
|---|
| "공", "일", "이", "삼", "사", "오", "육", "칠", "팔", "구" |
| "공일구칠", "일오팔삼", "이삼칠육", "삼팔육공", |
| "사구삼오", "오육일사", "육사이팔", "칠이사구", |
| "팔칠공일", "구공오이", "공육오삼", "일공이사", |
| "이사구공", "삼일칠팔", "사삼팔이", "오이육칠", |
| "육오삼일", "칠팔이구", "팔구사육", "구칠공오" |

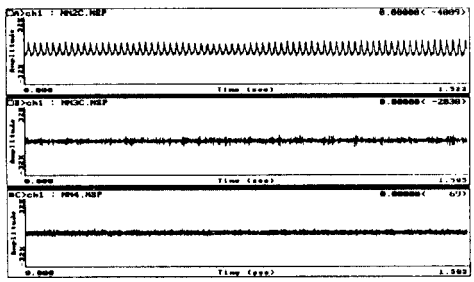


그림 5. 잡음(선풍기, 진공 청소기, 전화) 파형

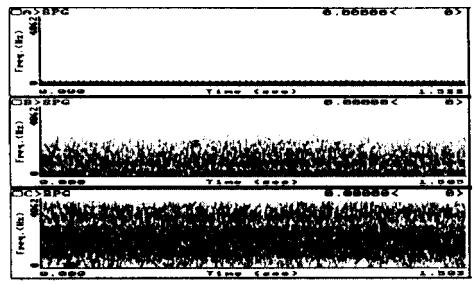


그림 6. 잡음 신호의 스펙트로 그램

## V. 실험 결과

### 5.1 결 과

본 논문에서는 작은 데이터 규모를 갖는 잡음 음성을 가지고 서로 다른 특징 추출 방법을 비교 하였다. 환경이 일치된 환경에서의 PLP 분석법은 상당히 효과적임을 알 수 있었다. 순시적인 신호대 잡음비 예측을 이용한 제안된 RASTA PLP 분석법은 잡음 환경에서 다른 분석법 보다 높은 인식율 보였다. 실험은 모두 화자 독립에 의해 이루어 졌다. 표 2는 깨끗한 환경과 잡음 환경의 인식 결과를 나타낸다. 표 3은 부가 잡음에 따른 각 분석법의 인식율을 SNR 15dB에서 나타낸 결과이다. 숫자음 인식결과 제안한 RASTA PLP 분석법이 변화된 환경과 잡음이 첨가된 데이터에서 가장 높은 인식율을 보였다.

표 2. 환경 변화에 따른 인식 결과

| 분석법 \ 환경 | [1]학습환경 | [2]잡음환경 |
|---|---|---|
| PLP | 96.0% | 83.0% |
| RASTA PLP | 94.0% | 92.0% |
| Proposed RASTA PLP | 96.0% | 94.0% |

[1]학습 환경 : 신경망 학습과 동일한 환경에서 수집한 데이터 (SNR 25dB 이상)

[2]잡음 환경 : 청소기 잡음등 잡음 환경에서 수집한 데이터 (SNR 10-15dB)

표 3. 부가잡음에 따른 인식 결과

| 분석법 \ 부가잡음 | 선풍기 잡음 | 진공청소기 잡음 | 전화기 잡음 |
|---|---|---|---|
| PLP | 84.0% | 80.0% | 76.0% |
| RASTA PLP | 88.0% | 86.0% | 80.0% |
| 제약한 RASTA PLP | 94.0% | 88.0% | 82.0% |

표 4는 J파라미터 값과 SNR에 따른 음성 인식 결과를 보여준다.

표 4. J값과 신호대 잡음비에 따른 인식결과

| SNR / J | 0dB－10dB | 10dB－15dB | 15dB－20dB | 20dB < |
|---|---|---|---|---|
| $10^{-10}$ | 70% | 74% | 78% | 79% |
| $10^{-9}$ | 80% | 82% | 91% | 90% |
| $10^{-8}$ | 76% | 90% | 92% | 94% |
| $10^{-7}$ | 72% | 86% | 93% | 94% |
| $10^{-6}$ | 68% | 84% | 88% | 96% |
| $10^{-5}$ | 60% | 84% | 86% | 96% |

구간별 J 파라미터 계산: 표 4에서 각 SNR에 따른 최고의 인식율을 보이는 값을 로지칼한 방법으로 해 당 SNR의 J값으로 구했다. (제안한 알고리즘)

표 5. 신호대 잡음비에 따른 구간별 J파라미터 값

| SNR | 0~10dB | 10~15dB | 15~20dB | 20dB < |
|---|---|---|---|---|
| J 파라미터 | $10^{-9}$ | $10^{-8}$ | $10^{-7}$ | $10^{-6}$ |

표 6은 학습에 참여하지 않은 공부터 구구구구까지 10000 단어중 임의로 50단어를 선택하여 실험한 결과이다.

표 6. 데이터 확장 실험 결과

| 환경 / 분석법 | 학습환경 (SNR > 25dB) | 잡음환경 (SNR < 15dB) |
|---|---|---|
| PLP | 82.0% | 68.0% |
| 제안한 RASTA PLP | 82.0% | 74.0% |

## Ⅵ. 결 론

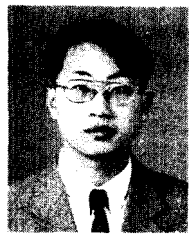
본 논문은 잡음 환경에 강한 음성의 특징 추출에 대하여 연구하였다. SNR을 이용한 제안한 RASTA PLP분석법은 잡음에 의한 왜곡을 감소시킬 뿐만 아니라 환경 변화에도 잘 적응하여 RASTA 처리를 하지않은 PLP 보다 높은 인식율을 보였다. 환경 변화에 따른 노이즈 특성 변화는 많은 음성처리 알고리즘의 성능을 감소시키나 본 논문에서 사용된 순시적인 SNR예측 알고리즘은 계산이 덜 복잡하고 노이즈 특성이 변하는 문제에 잘 대처할 수 있다. 그러나 이 알고리즘도 음성이 없는 구간에서는 다소 왜곡이 있다. 환경이 불일치한 경우의 음성 인식기의 성능이 감소함을 볼 수 있었다. 그러나 우리는 RASTA 필터링과 순시적인 신호대 잡음비 예측을 통해 이 영향을 성공적으로 줄였다. 앞으로의 연구 계획은 특징추출에서 시간적 정보를 더 포함하고, 부가된 환경 노이즈를 억제하며, 더 많은 학습 세트(성, 발음, 액센트, ...)를 구성하여 실험할 계획이다. 또한 실험 데이터를 잡음환경의 범위를 확대하여 수집하여 실험하고, 잡음 환경에 적절하게 동작하여 실제 생활에 응용이 가능하리라고 판단된다.

## 참 고 문 헌


1. B.A. Hanson, T.H. Applebaum, "Subband or Cepstral domain Filtering for Recognotion of LOMBARD and Channel-Distorted Speech", Proc. *ICASSP* Ⅱ.79～Ⅱ.82 1993.
2. H. Hermansky, N. Mogan, H. Hirsh, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing", *ICASSP*, pp Ⅱ.83-Ⅱ.86, 1993.
3. H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech", *J. Acoust. Soc. Am.*, pp. 1738～1752, 1990.
4. Martin. R, "An efficient algorithm to estimate the intantaneous SNR of speech signals", *EUROSPEECH*, pp. 1093-1096 1993.
5. 홍기원, 김선일, 송도선, 김석동, 이행세, "인공 신경망과 MARKOV 모델을 이용한 한국어 단어 인식에 관한 연구", *대한전자공학회 하계학술발표 논문집*, vol. 18, no. 1, pp. 629-632, June 1995.
6. 이 행세, *음성인식*, 청문각, 1996.



▲배 현 권(Hyun Kwon Bae)



1994년 2월: 경기대학교 전자공학과 (공학사)

1993년 12월~1995년 12월: (주)진성 아이시스 부설연구소 연구원

1996년 3월~현재: 아주대학교 전자공학과 석사과정

※주관심분야: 음성인식, 패턴인식, 디지털 신호처리

▲오 문 식(Moon Sik Oh)

1997년 2월: 명지대학교 전자공학과 (공학사)

1997년 3월~현재: 아주대학교 전자공학과 석사과정

※주관심분야: 음성인식, 패턴인식, 영상 신호처리

▲이 행 세(Haing Sei Lee)

현재: 아주대학교 교수·전자전기공학부