스펙트럴 정보를 이용한 잡음환경에서의 음성인식

이광석*, 김현주**

요약

본 연구는 잡음을 포함한 음성 환경에서의 음성인식 성능의 개선방안에 관한 것이다. 우리는 음성인식의 성능평가 고찰로부터 잡음부가 음성의 스펙트럴 포락으로부터 곡들의 스펙트럴 탈락 및 복원이 보다 더 효과적임을 알 수 있었다. 본 연구에서는 평균화된 스펙트럴 포락을 모음 스펙트럴 으로 부터 추출하여 곡의 강조에 사용하였으며 이들의 특징은 낮은 주파수 영역에서의 모음 스펙트럴 정보는 강조되어지고 자음으로부터 얻은 스펙트럼은 변하지 않는다는 것이다. 한편, 강조계수는 켑스트럴 영역에서 변하므로 이를 이용하여 잡음 있는 숫자음성 인식에 적용하였으며 인식결과가 양호하게 개선됨을 알 수 있었다.

Speech Recognition using Spectral Information under Noisy Environments

Gwang-Seok Lee*, Hyun-Ju Kim**

ABSTRACT

This research is concerned for improving the result of speech recognition under the noisy speech. We knew that spectral subtraction and recovery of valleys in spectral envelope obtained from noisy speech are more effective for the improvement of the recognition. In this research, the averaged spectral envelope obtained from vowel spectrums are used for the emphasis of valleys. The vocalic spectral information at lower frequency range is emphasized and the spectrum obtained from consonants is not changed. In simulation, the emphasis coefficients are varied on cepstral domain. This method is used for the recognition of noisy digits and is improved well.

Key Words : Speech Recognition, Noisy Speech Processing, Spectral Envelope, Emphasis Coefficient

* 진주산업대학교 전자공학과(✉kslee@jinju.ac.kr)

** 진주산업대학교 컴퓨터공학과

· 제1저자(First Author) : 이광석 · 교신저자(Correspondent Author) : 김현주

· 접수일(2010년 9월 20일), 수정일(1차 : 2010년 10월 13일), 게재확정일(2010년 10월 19일)

Ⅰ. 서 론

잡음환경에서의 음성정보처리 및 음성인식은 실용화시 중요한 문제로 인식되고 있으며 다양한 연구가 진행되고 있다[1][2][3]. 이는 입력파형에 대하여 직접적인 잡음처리를 행하는 방법과 잡음중첩 음성을 분석한 후, 그 분석 파라미터에 대하여 잡음처리를 행하는 방법으로 나눌 수 있다. 대표적으로 스펙트럴 영역에서의 잡음처리를 들 수 있으며 이는 잡음환경의 음성인식에서 스펙트럴 포락변형으로 대처한 대표적 방법으로서 보통은 스펙트럴 탈락이 생각되지만 이 경우, 발생한 잡음 스펙트럴을 완전히 추정해 둘 필요가 있으며 이를 위하여 일차함수를 사용한 입력음성 스펙트럴의 낮은 레벨 제한을 제안하고 이것을 준정상 유색잡음 환경하의 음성인식에 적용하였다[1]-[5]

그 결과, 의사 스펙트럴 탈락 가능한 것, 중첩잡음 스펙트럴의 스펙트럴 경사의 정도 및 제한하는 일차함수의 경사사이에서 10dB정도의 차이가 생기는 것은 인식에 대한 악영향이 적다는 것을 나타낸다. 잡음중첩 유성음 스펙트럴에서 스펙트럴 포락으로 곡 부분이 탈락하는 것에 착안하여 규칙에 의해 탈락한 곡 부분의 회복방법을 숫자 음성인식에 적용하여 그 유효성에 대하여 나타내었다. 이 고주파 영역의 정보는 통상 자음부의 식별에서 보다 중요하다는 것을 고려하면 저주파 영역의 강조도를 음소마다 변화시키는 것이 유효하며 또한 고 레벨의 스펙트럴 요소변화의 보정을 위하여 곡이나 경사성분의 강조도를 조정하여 시행하는 것이 중요한 요소라는 것을 알 수 있었다.

스펙트럴 포락의 저주파 영역을 강조키 위한 방법으로써 스펙트럴 포락의 주파수 축의 비직선 변형에 의한 평가를 들 수 있다. 이것은 최소위상의 스펙트럴 포락과 연관된 파라미터에 재귀식으로 계산하고 파라미터 분석에 적응하기가 쉬우며 적당한 주파수축도 산출되고 있다. 음성의 각 시점에 있어서 스펙트럴을 모음평균 스펙트럴로써 강조하는 경우, 그 곡 정도를 나타내는 요소를 캡스트럼의 Quefrency상에서 분리하고 각각 효과적인 강조법을 검토하였다. 이러한 파라미터에 의한 강조법을 잡음 환경하에서의 숫자 음성인식에 적용하고 인식 파라미터로서의 유효성을 검증하였다. 또한, 종래의 스펙트럴의 규칙적인 곡부가 병용으로 Over-emphasis에 대응하기 위하여 거리의 가중치 부가 방법에 대해서도 함께 검토하였다.

Ⅱ. 잡음 부가된 스펙트럴 포락의 곡에 의한 스펙트럴 보정

종래의 스펙트럴 레벨 제한과 규칙에 의한 곡의 부가에 대하여 살펴보면 잡음중첩 유성음의 경우, 스펙트럴 포락의 저 레벨부만의 레벨 상승변화를 볼 수 있으며 이 현상이 일어나는 주파수 범위는 1~2 ㎑부근 및 제2포만트 주파수 이후에 일어나는 것을 확인할 수 있다.

한편, 강조계수에 의한 보정과 의미 있는 규칙을 사용하여 유성부 스펙트럴 저 레벨에 곡을 부가하고 이 결락을 보정하는 것을 생각할 수 있으며 다음과 같다. 유성부 스펙트럴 포락에 대하여 우선 (1)식과 같이 일차함수 th(k)를 사용하여 간이 스펙트럴 탈락을 행하고 이를 스펙트럴 저 레벨 제한으로 부르고 있다. 여기서, k는 이산주파수를, N는 표본수를 각각 나타낸다.

$$S'(k) = \begin{cases} S(k), & (S(k) > th(k)) \\ th(k), & (S(k) \le th(k)) \end{cases}$$
$$th(k) = TH1 + 2(TH2-TH1)k/N \quad (1)$$

그리고 규칙에 의한 곡의 부가변형을 행하며 데이터의 표본화 주파수 10㎑일 때 다음과 같이 나타낸다.

$$THL(f) = p(f1) + \{W(f3)-p(f1)\}\{(f - f1)/(f3 - f1)\}$$

$$THL(f) = p(f2) + \{W(f3)-p(f2)\}\{(f2 - f)/(f3 - f2)\}$$
$$p(f) = TH1 + (f /5000)(TH2 - TH1) \quad (2)$$

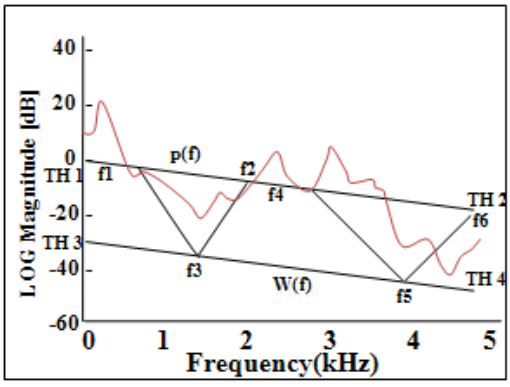


그림 1. 규칙에 의한 스펙트럴 포락에의 곡의 부가
Fig. 1 Addition of valley to spectral envelope by rule

단, 하나의 곡에 대하여 THL은 곡의 좌측을, THR은 곡의 우측을 나타내는 직선 함수이며 p(f)는 잡음 레벨을 의사하는 직선함수로서 (1)식의 th(k)와 동일 함수이다. 또한 W(f)는 곡의 깊이를 의사하는 직선함수이다. 제1곡에 대해서는 식의 f1, f2는 1.2㎑이후의 주파수로써 원 스펙트럴 포락과 P(f)의 교점 주파수이며 f3는 그 중심점으로 그림 1과 같이 곡을 부가한다. 제2곡에 대해서는 f2이후의 주파수로 원 스펙트럴 포락이 P(f)를 위에서 아래로 지나는 주파수를 새로운 f1으로 하여 제1곡과 같은 형식으로 부여한다. 이와 같이 부여한 곡의 수는 저주파 영역부터 순서대로 두 개로 한정하고 있으며 W(f)의 임계치는 스펙트럴 포락 제한・곡용으로 각각 두 개씩 설정하고 있다.

III. 모음 스펙트럴을 이용한 유성부 스펙트럴 포락의 경사 강조

앞에서 유성음, 특히 유성 모음성분의 스펙트럴의 저역부 강조방법에 대하여 언급했지만 잡음 중첩음성에서 모음성 음소의 스펙트럴 포락의 곡 부분이 묻힌 스펙트럴 포락 전체가 평탄화 되기 때문에 이것을 보정하는 의미로서 모음성 스펙트럴 포락에 대하여 단모음의 스펙트럴 포락을 사용하여 다음과 같이 스펙트럴 포락 경사의 강조를 행하는 것을 고려한다. i차의 캡스트럼을 Ci, 모음평균 캡스트럼을 νi로 할 때 (3)식과 같은 변형을 유성부에만 행하는 것으로 하며 β는 강조계수이다.

$$C'_i = C_i + \beta\nu_i \quad (3)$$

여기서는 모음 캡스트럼으로 시뮬레이션에 사용한 단어음성 화자와 별개의 남성화자가 발성한 5모음의 중심부 스펙트럴 포락에서 구한 것을 사용한다. 이 포락의 예를 그림2에 나타내었다.

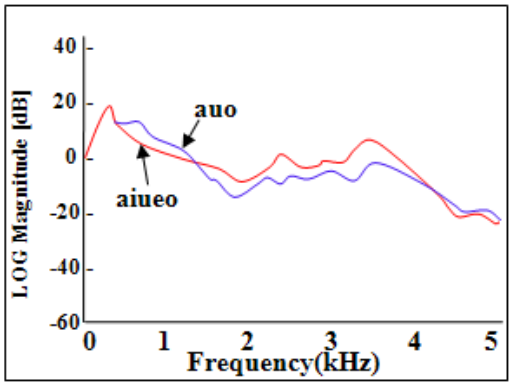


그림 2. /a/, /u/, /o/ 및 5자음의 스펙트럴 포락
Fig. 2 Spectral envelope of /a/, /u/, /o/ and 5 vowels

양자의 스펙트럴 포락형상에는 본질적인 차는 없지만, 여기서는 스펙트럴 포락에서의 경사와 곡 부분을 강조할 수 있는 /auo/로부터 얻어진 포락을 사용하고 있다. 이것을 의사 모음성 음소필터로 생각하고 스펙트럴 강조에 사용한다. 식(3)에 의해 변화시킨 스펙트럴 포락변화의 예를 그림 3에 나타내었으며 남성화자 모음 /i/에 대하여 그림2의 /auo/포락을 사용하

여 강조한 것으로서 스펙트럴의 곡의 강조와 스펙트럴 평탄화의 보정에 이용 가능함을 알 수 있다. 여기서, 그림 2의 포락은 캡스트럼의 특징에 의해서 Quefrency영역에서 분리하는 것으로 생각된다. 모음 캡스트럼을 보면 이전의 모음 연쇄 분할 시뮬레이션에 의해서 스펙트럴 포락의 대표적 경사를 표시한 Quefrency저차부와 포만트 미세변화를 표시한 Quefrency고차부에서 분리된다.

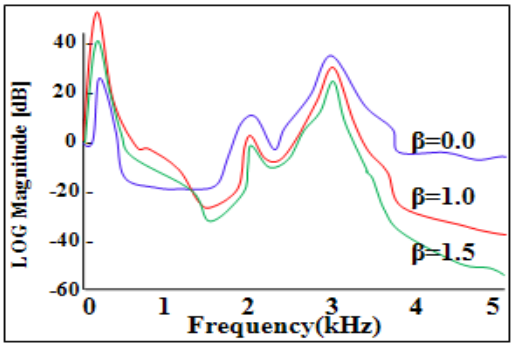
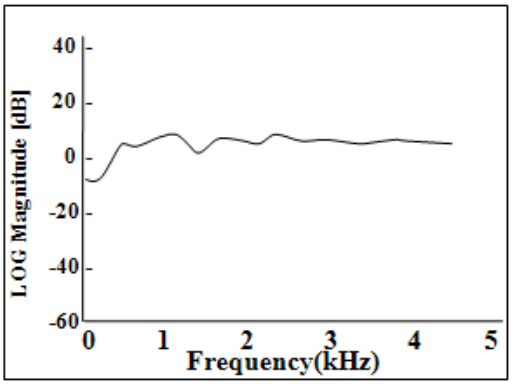
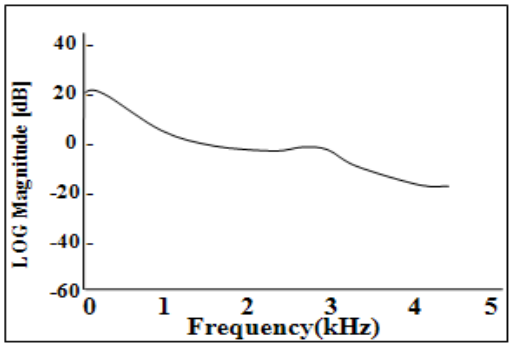


그림 3. /i/에 대한 모음/auo/의 평균 스펙트럴 강조
Fig. 3 Emphasis of average spectral of vowels (/auo/) to /i/

그림 4는 그림 2의 포락을 Quefrency저차부와 고차부에서 분리한 것과 각각의 포락을 표시하지만 이러한 특징을 표시하고 있는 것을 알 수 있다.



(a) 1st~3nd order



(b) 4th~12th order

그림 4. Quefrency분리 후의 모음/auo/의 평균포락
Fig. 4 Average envelope of vowels /auo/ after quefrency separation

여기서 식(3)의 모음 캡스트럼에 의한 경사 강조계수를 캡스트럼 영역에서 이 두 종류를 분리하고 강조계수를 변화시키는 것으로 생각할 수 있다.

$$e_i = \begin{cases} C_i + \beta_1 v_i & ,(\ 1 \le I \le m\ ) \\ C_i + \beta_2 v_i & ,(\ m+1 \le i \le n\ ) \end{cases} \quad (4)$$

(4)식에서, β1은 대표적 포락 경사의 강조, β2는 미세한 포락의 강조, 포락 피크선의 강조를 표시한 것으로 생각할 수 있으며, 이 경우에는 유성부의 스펙트럴 포락의 곡부분과 스펙트럴 포락 전체의 경사도가 강조됨이 알려져 있다.

Ⅳ. 시뮬레이션 및 고찰

남성화자 10명이 5회 발성한 10개의 숫자음성에 2종의 유색잡음을 더하여 시뮬레이션에 사용하였으며 각각은 경사 10dB, 자동차는 경사 30dB이고, 피크로 표기된 잡음쪽이 백색잡음에 가까운 성질을 가지고 있다. 이것은 프레임 주기 10㎳에서 멜-캡스트럼 분석이 되며 스펙트럴 변형은 앞에서 캡스트럼에 의한 경

사 강조가 일어난 후 정합되므로 재귀식에 의해 멜-주파수축으로 주파수 변환된다. 포락 변환일 경우 유성·무성 판단은 무잡음 데이터의 캡스트럼 저차부의 합으로 판단하고 있다. 시뮬레이션에서는 5프레임 끝점-개방 DP매칭에 의해 불특정 화자인식에 적용하였으며 표준패턴은 무잡음 음성을 사용하였다. 시뮬레이션 결과의 예로써 표1은 식(3)에 의한 일정한 강조계수에 의한 강조를 행한 경우의 인식결과의 예를 보이고 있으며 강조계수 β는 1.0으로 된다.

표 1. 모음 평균 포락에서 동일한 강조에 의한 인식결과 (β=1.0)
Table 1 Recognition results by equal emphasis in vowel average envelope (β=1.0)

Emphasis Order		Vehicle		Peak	
Start	End	0dB	10dB	0dB	10dB
1	12	36.2	55.5	28.8	47.5
2	12	29.5	55.9	21.5	40.6
1	10	36.3	54.3	28.2	46.6
2	10	30.2	55.6	21.8	38.5
1	8	35.3	56.3	26.4	45.2
2	8	30.6	55.5	20.6	34.5
1	5	36.5	55.8	24.3	46.2
2	5	25.4	54.5	17.5	32.4
1	3	35.3	56.2	27.3	46.4
1	2	36.4	56.6	26.3	46.8

강조차수는 매칭차수와 같은 방법으로의 적용한 경우에 좋은 결과를 얻었으며 대표적 경사보정을 행한 임의 캡스트럼 일차항이 중요함을 알 수 있었다. 고주파영역의 잡음성분이 감소한 경우 이 효과의 정도는 작게 되는 경향이 있다. 표2에서 식(2)에 의한 원래의 스펙트럴 유성부에 2개의 곡을 부가하고 표1과 같은 조건의 강조를 행한 결과를 표시하였다. 표3에서는 표1과 같은 강조 시뮬레이션에 대하여 식(4)에서 강조계수를 Quefrency상에서 변화시킨 경우의 시뮬레이션 결과를 표시하였다.

시뮬레이션은 잡음 스펙트럴 포락에 대하여 곡을 부가하지 않는 경우에 대해서 행한 것이고 강조는 앞의 시뮬레이션과 같게 미리 구한 유성부에 대하여 행하며 이러한 경우, 저차부의 강조도를 강화시키고 있다.

표 2. 곡 부가를 함께 사용한 모음 평균 포락에서 동일 강조에 의한 인식결과 (β=1.0)
Table 2 Recognition results by equal emphasis in vowel average envelope together use of valley addition(β=1.0)

(a) TH1=0dB, TH2=-10dB, TH3=-30dB, TH4=-40dB

Emphasis Order		Vehicle		Peak	
Start	End	0dB	10dB	0dB	10dB
1	12	33.5	50.2	33.4	60.7
2	12	29.6	44.5	25.5	47.2
1	10	35.2	49.4	30.5	57.5
2	10	27.8	43.7	25.2	46.2
1	8	35.3	48.2	31.6	58.7
2	8	27.9	40.5	23.2	44.4
1	5	33.4	49.6	31.8	59.3

(b) TH1=0dB, TH2=-20dB, TH3=-30dB, TH4=-40dB

Emphasis Order		Vehicle		Peak	
start	end	0dB	10dB	0dB	10dB
1	12	35.2	65.5	30.1	56.3
2	12	31.5	66.5	22.3	44.2
1	10	35.5	65.4	29.5	55.3
2	10	30.6	65.4	23.6	43.2
1	8	35.8	65.2	29.4	55.5
2	5	30.4	65.3	21.2	40.4
1	3	36.3	64.6	27.8	56.6

표 3. 모음포락에 의한 강조에서 강조계수를 변화시킨 경우의 인식결과 : 정합차수 : 1차~12차
Table 3 Recognition results In case of varying emphasis coefficient in emphasis by vowel envelope : matching order : 1st~12th

(a) m=3, n=12

Emphasis Coeff.		Vehicle		Peak	
Low	High	0dB	10dB	0dB	10dB
1.2	1.0	37.8	55.6	33.1	47.4
1.5	1.0	36.9	54.8	34.6	46.2
1.8	1.0	36.1	53.5	34.9	46.3
2.0	1.0	35.5	51.9	36.2	46.1

(b) m=5, n=12

Emphasis Coeff.		Vehicle		Peak	
Low	High	0dB	10dB	0dB	10dB
1.2	1.0	35.5	55.2	32.6	47.8
1.5	1.0	34.8	53.5	35.2	47.2
1.8	1.0	34.2	52.4	35.6	46.6
2.0	1.0	35.6	48.3	36.3	45.1

결과적으로 나쁜 환경의 경우 인식률이 향상되고 있지만, S/N비가 나쁘지 않은 경우, 또한 잡음 환경이 백색성에 가까운 경우, Over-emphasis가 되는 것을 생각할 수 있다. 이 때문에 DP매칭에서 유클리드 거리를 계산할 때 전 프레임에 관한 식(5)의 가중치를 부가한 거리를 사용하여 이점의 보정을 행한다.

$$d^2 = \sum_{i=1}^{j}\{\gamma_1(e'_i - e_i)\}^2 + \sum_{i=j+1}^{n}\{\gamma_2(e'_i - e_i)\}^2 \qquad (5)$$

식(5)를 사용한 경우의 시뮬레이션 결과의 예를 표 4에 나타내었다. 여기에서, γ_2의 값은 1.0으로 고정하고 γ_1의 값의 변화에 의해서 캡스트럼 저차부의 가중치를 변화시키고 있으며 이 경우, $\gamma_1 < 1.0$ 이기 때문에 저차부 거리를 저감시키고 있는 것으로 된다. 이와 같이 저차에 작은 가중치를 곱함으로써 전자의 문제가 해결된 경우를 표시하였다. 특히 백색성에 가까운 잡음 및 잡음조건이 그다지 나쁘지 않는 경우에 개선됨을 알 수 있으나 이 방법은 조건에 의해 최적치가 민감하기 때문에 좀 더 검토가 필요하다.

표 4. 정합계수 m=3, n=12, β_1=1.5, β_2=1.0, γ_2=1.0에 의한 인식결과
Table 4 Recognition results by matching coefficient : m=3, n=12, β_1=1.5, β_2=1.0, γ_2=1.0

(a) 곡을 부가하지 않은 경우

Matching Coeff.		Vehicle		Peak	
Order	γ_1	0dB	10dB	0dB	10dB
1-3	0.7	36.6	56.3	35.6	48.3
1-3	0.8	36.4	56.1	35.3	47.6
1-3	0.9	35.4	55.4	34.5	47.1
1-1	0.7	34.2	54.2	34.1	49.2
1-1	0.8	35.3	53.8	35.2	48.7
1-1	0.9	35.6	53.1	34.2	48.5

(b) 곡을 부가한 경우

Matching Coeff.		Vehicle		Peak	
Order	γ_1	0dB	10dB	0dB	10dB
1-3	0.7	35.6	51.2	36.6	58.3
1-3	0.8	35.3	50.6	36.4	58.1
1-3	0.9	34.5	50.1	36.1	57.4
1-1	0.7	33.4	51.3	35.6	61.8
1-1	0.8	33.9	50.5	36.5	62.4
1-1	0.9	34.3	49.6	36.2	59.7

: TH1=0dB, TH2=-10dB, TH3=-30dB, TH4=-40dB

V. 결 론

본 연구에서는 모음을 대표로 하는 유성부와 무성부에 대하여 저역 스펙트럴의 중요도의 차이에 대하여 검토하였다. 또한, 잡음환경에서 음성의 유성부 스펙트럴 포락의 변형에 주목한 곡의 보정법을 위하여 모음 스펙트럴 포락을 강조필터로써 사용하였다. 스펙트럴 포락의 떨어진 곡의 강조 및 평탄화한 기울기 강조법에 관하여 제안하고 그 유효성을 검토하였다. 그 결과, 잡음환경이나 정합조건에서 다소 차이가 나지만, 특히 잡음이 존재하는 환경하에 잡음 스펙트럴이 평탄한 경우에 유효함을 알 수 있었다. 또한 Over-emphasis 보정을 위해 가중치를 부가한 거리에 대하여 시뮬레이션을 시행하고 검증하고 그 실효성을 알 수 있었다.

향후 주파수변환 계수변화의 크기 판단 및 각각의 조건에 관한 최적 변형방법에서의 검토 및 거리 척도를 포함한 제안 방식에 의한 스펙트럴 변형을 사용한 경우의 매칭방식의 검토와 연구가 계속하여 진행하여야 할 것으로 생각된다.

참고문헌

[1] Davis S. B. and Mermelstein P. "Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoust., Speech Signal Processing, 28, 357-366, 1980.

[2] J-W. Hung, J-L. Shen and L-S Lee, "New Approaches for Domain Transformation and Parameter Combination for Improved Accuracy in Parallel Model Combination (PMC) Techniques", IEEE Trans. Speech and Audio Processing, 9(8), 842-855, 2001.

[3] Kamath, S., Loizou, P. (2002). " A Multiband Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise", Proc. IEEE Int. Conf. Acoust., Speech Signal Processing, pp.101-111.

[4] Kim, N.S. (2003). "Speech Recognition under Noisy Environments", 나 Telecommunications Review, pp. 650-661.

[5] ESTI Drafr Standard Doc. Speech Processing, (2000). Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithms; Compression Algorithm, ESTI ES 202 108 V1.1.2 (2000-04), http://pda.eti.org/pda/queryform.asp, April.

이광석(Gwang-Seok Lee)

1983년 동아대학교 전자공학과(공학사)
1985년 동아대학교 대학원 전자공학과 (공학석사)
1992년 동아대학교 대학원 전자공학과 (공학박사)
1995년~현재 국립진주산업대학교 전자공학과 교수
※ 관심분야 : 음성신호처리 및 인식, 퍼지 및 신경회로망, Biometrics, 지능화 기술

김현주(Hyun-Joo Kim)

1988년 경상대학교 컴퓨터과학과(이학사)
1990년 숭실대학교 대학원 전자계산학과 (공학석사)
2000년 경상대학교 대학원 컴퓨터과학과 (이학박사)
2002년~현재 국립진주산업대학교 컴퓨터공학과 부교수
※ 관심분야 : 정보검색, XML, 컴퓨터교육, 데이타마이닝