퍼지 및 지능시스템학회 논문지 2003, Vol. 13, No. 1, pp. 45-50

음향적 요소분석과 DRNN을 이용한 음성신호의 감성 인식

Analyzing the Acoustic Elements and Emotion Recognition from Speech Signal Based on DRNN

심귀보* · 박창현* · 주영훈**

Kwee-Bo Sim* · Chang-Hyun Park* · Young-Hoon Joo**

* 중앙대학교 전자전기공학부

** 군산대학교 전자정보공학부

요 약

최근 인간형 로봇에 대한 개발이 괄목할 만한 성장을 이루고 있고, 친근한 로봇의 개발에 중요한 역할을 담당하는 것으로써 감성/감정의 인식이 필수적이라는 인식이 확산되고 있다. 본 논문은 음성의 감정인식에 있어 가장 큰 부분을 차지하는 피치의 패턴을 인식하여 감정을 분류/인식하는 시뮬레이터의 개발과 시뮬레이션 결과를 나타낸다. 또한, 피치뿐 아니라 음향학적으로 날카로움, 낮음 등의 요소를 분류의 기준으로 포함시켜서 좀더 신뢰성 있는 인식을 할 수 있음을 보인다. 주파수와 음성의 다양한 분석을 통하여, 음향적 요소와 감성의 상관관계에 대한 분석이 선행되어야 하므로, 본 논문은 사람들의 음성을 녹취하여 분석하였다. 시뮬레이터의 내부 구조로는 음성으로부터 피치를 추출하는 부분과 피치의 패턴을 학습시키는 DRNN 부분으로 이루어져 있다.

ABSTRACT

Recently, robots technique has been developed remarkably. Emotion recognition is necessary to make an intimate robot. This paper shows the simulator and simulation result which recognize or classify emotions by learning pitch pattern. Also, because the pitch is not sufficient for recognizing emotion, we added acoustic elements. For that reason, we analyze the relation between emotion and acoustic elements. The simulator is composed of the DRNN(Dynamic Recurrent Neural Network), Feature extraction. DRNN is a learning algorithm for pitch pattern.

Key Words : 피치(Pitch), 포만트 주파수(Formant Frequency), 음질, DRNN

1. 서 론

최근에 개발, 발표 되고 있는 로봇들은 두발로 걷고, 춤을 추는 등 인간의 행동과 매우 유사한 동작을 할 수 있다. 물론, 아직 영화나 만화에서 꿈꿔오던 것처럼 실제 인간과 구분이 힘들 정도로 동작할 수 있는 기술까지는 부족하지만, 현 시점에서는 그 정도의 움직임을 보이는 것만으로도 매우 놀라운 발전임은 분명하다. 그러나, 단순한 움직임만을 구현하는 것은 인간의 보조자로써 한계가 있다. 기계의 업무 영역을 넓히기 위해서는 좀더 인간과 유사해져야 한다. 이때 필요한 것이 감성의 인식이다. 인간의 감성을 인식하게 됨으로써 좀더 유연한 방법으로 인간에게 도움을 줄 수 있을 것이다.

예를 들면, 어떤 사람이 로봇에게 음악을 틀라고 명령을 내렸을 때, 로봇이 사람의 감성을 인식한다면, 적절한 음악의 목록을 제시하고 명령을 수행할 수 있을 것이다. 그리고, 로봇 이외에도 게임 등의 소프트웨어에도 적용하여 더욱 재미있는 제품들이 만들어 질 수 있을 것이다. 감성 인식은 얼굴 표정인식과 음성으로부터의 인식, 두 가지 접근 방법이 있다. 본 논문에서는 두 방법 중 음성으로부터의 인식을 연구하였다. 기존의 논문은 피치를 주요 특징으로 이용하였다. 피치는 물리적으로 음의 높낮이를 뜻하는 것으로써 사람의 감성 정보가 가장 많이 포함되어 있다. 그런 이유로 많은 연구자들이 피치를 주요 특징으로 사용한다. L. S. Chen은 피치의 통계적 특성과 에너지 궤적 등을 이용하였다. 그는 음성만으로 감성인식을 하는 것은 한계가 있고, 화상인식과 함께 하는 경우에 그 성능이 향상된다는 것을 주장하였다. Joy Nicholson은 음성 데이타를 기본으로 하였으나 그도 역시 피치와 음성의 Power를 특징으로 이용하여 감성인식을 시도하였다. 또한, 학습방법으로는 신경망을 기반으로 하였다. 그의 경우에는 Sub-neural network(MLP)들을 감정의 수만큼 배치하고 각 Sub-neural network의 출력 값들을 Decision Logic에 입력시켜 최종 출력 값을 얻어내었다. 본 논문은 피치의 패턴을 감성 인식의 주요 요소로써 사용할 것이고, 부가적으로 각 감성과 음향적 요소의 관계를 실험을 통해 알아본다. 즉, 4가지 감정(평상, 화, 웃음, 놀라움) 상태에서 발화된 문장에 대한 분석을 통하여 각 감성상태에서의 날카로움,

접수일자 : 2002년 11월 9일
완료일자 : 2003년 2월 1일
본 연구는 산업자원부의 2000년도 차세대신기술개발사업인 『수퍼지능칩 및 응용기술개발』 과제의 제5세부과제인 「Autonomous Family Machine(AFM) 요소기술개발(N09-A08-4301-05)」의 위탁연구로 이루어졌으며, 산업자원부의 연구비지원에 감사드립니다.

낮음, 크기, 통계적 특성 등의 값을 찾아내는 것이다. 그리고, 이러한 특징들을 이용하여 분류, 학습의 과정을 거쳐 인식을 한다. 학습과정에서는 음성신호의 특성에 맞춰 DRNN (Dynamic Recurrent Neural Network) 을 사용한다. 인간의 두뇌는 단순하고 정적인 시스템이 아니라 고차 비선형 동적 시스템이다. 그러므로 이러한 두뇌를 모델링하기위해서는 복잡한 역동성을 구현하고 내부 상태를 저장할 수 있는 시스템이 필요한데, 이러한 특성을 갖춘 것이 바로 DRNN이다. 이러한 학습과정을 거쳐 학습된 신경망으로 여러 화자와 여러 내용의 음성을 테스트하였다.

2. 본 론

2.1 음향의 감성정보

소리의 정보전달에는 자료의 전달뿐 아니라 정서적인 정보 또한 포함하고 있다. 동일한 정보를 전달하더라도 부드러운 목소리로 전달하는 경우와 불쾌한 목소리로 전달하는 경우의 정보전달 효과는 다르다. 소리 신호는 푸리에의 이론에 의하면 여러 주파수들의 조합으로 이루어져있으므로 갖가지 감성을 갖는 소리들을 주파수 영역에서 분석하여 분류할 수 있다. 목소리의 기본음(Fundamental Frequency)은 125Hz~250Hz 이고, 목소리의 힘을 주는 음역대는 350Hz에서 2000Hz 다. 즉, 에너지가 가장 많은 부분이다.

표 1. 주파수 대역에 따른 청각적 느낌.
Table 1. The relations of frequency and feeling

주파수	청각적 느낌
125Hz~500Hz	증폭하면 묵직함이 생기고, 줄이면 굉장히 약한 목소리가 됨
2000Hz~5000Hz	증폭하면 목소리의 명확도가 높아짐
4000Hz~8000Hz	치찰음, S, SH, CH, C 등을 말할 때의 듣기 싫은 소리가 나옴. 줄이게 되면 목소리의 깨끗함도 감소 됨.
8000Hz 이상	입에서 나오는 공기소리

위의 표는 주파수 대역에 따라 다른 청각적 느낌을 나타낸다

2.1.1 음향 요소 분석

본 논문은 음향요소를 날카로움, 저음, 굵음, 가늚, 큼, 작음 의 6가지로 정의한다. 6가지 요소의 분석을 위하여 모음' 아'에 대한 실험을 하였다. 표 2 에서 F : Formant , Mag : Magnitude , Int : Intensity, NU : Non Uniform , M1: Man 1 을 의미한다.

표 2, 3은 여러 실험 데이터 중 대표적인 결과이다. 위 결과에 따르면 음의 높낮이는 Pitch로 분명한 구분이 가능하다. 그리고, 날카로운 소리의 경우는 3,4,5 Formant가 순간순간 변하는 것을 알 수 있다. 날카로운 소리와 저음의 경우를 비교해보면, 배에서부터 나온 소리의 경우 Uniform한 Formant의 분포를 보이나, 날카로운 소리 중 머리로부터 울리는 경우에는 3,4,5 Formant에서 Non Uniform한 특성을 보인다. 이 경우에는 소리가 맑지 않다. 그리고, 125 - 500Hz는 풍부한 느낌의 주파수가 분포한다. 또한, 1F가 500Hz 주변에서 많이 분포하는 경우에 더 굵은 목소리로 들린다. 표 2에서 저음과 중 저음 부분의 1F,2F가 같은 대역에 분포하는 경우는 각각 다른 대역에 분포하는 경우보다 에너지의 크기가 더 크므로 더욱 굵은 소리를 낸다. 반대로 500Hz에서 멀수록 가는 소리이다.

표 2. M1에 대한 음향적 분석
Table 2. Acoustical analysis for M1

	M1_평	M1_날카로움	M1_저음	M1_중저음
1F(Hz)	824	800	600	650
2F(Hz)	1100	1200		
3F(Hz)	3000	NU	2800	2780
4F(Hz)	3480		3300	3400
5F(Hz)	4570		NU	4300
Mag	0.8	1.65	0.6	1.4
Int(db)	78db	88db	75db	80db
Pitch	134Hz	370Hz	109Hz	130Hz

표 3. M2에 대한 음향적 분석
Table 3. Acoustical analysis for M2

	M2_평	M2_날카로움	M2_저음	M2_중저음
1F(Hz)	812	700	630	743
2F(Hz)	1210	1200	1100	1160
3F(Hz)	2760	2700	2500	2600
4F(Hz)	3700	NU	3400	3400
5F(Hz)	NU		없음	3600
Mag	1.6	1.6	0.4	0.8
Int(db)	83	87	65	74
Pitch	117	290	96	103

표 4. 음성 파형의 분산과 스펙트럼 분석
Table 4. Variance and spectrum analysis

[아]	낮음		평서		높음	
	분산	S	분산	S	분산	S
A	13356	6	13494	20	14386	35
B	13298	4	13434	26	14503	42
C	13298	1	13540	23	14518	55

(S는 기준치를 넘은 스펙트럼 포인트 개수.)

표 4는 성인 남성 3명에 대해 [아]를 3가지 높이로 발화한 경우의 음성 파형의 분산과 스펙트럼 분석해본 결과이다. 보통 음성을 격하게 발화하는 경우 음성 파형 또한 격해지는 모양을 보이는 특성으로부터 분산의 비교를 하였다. 표에서 보는 바와 같이, 날카롭게(높게) 발화하는 경우 분산이 높아지고, 낮은 음일 수록 분산이 낮아지는 성향을 볼 수 있다. 또한, S 는 주파수 영역에서의 분석을 단순화 한 것으로써, 수많은 주파수 성분을 전부 언급할 수 없으므로, 스펙트럼의 크기가 일정 기준치를 넘은 경우의 개수를 기록한 것으로써, 표에는 나와 있지 않지만, 낮은 음인 경우는 S가 낮을 뿐 아

니라 고주파영역에는 S가 전혀 없었다. 반대로, 높은 음인 경우는 S가 높고(큰 에너지대가 많다.) 고주파 성분 또한 전체 에너지의 10~20%가량 분포하는 것을 볼 수 있다. 앞으로 이러한 분석을 바탕으로 날카롭게(높게) 발화하는 감정의 표현이나 낮게 발화하는 감정의 표현에 대한 구분에 대한 신뢰성을 더할 수 있다.

2.1.2 감정과 파라메터의 관계

■ 평서형:
Pitch Contour가 평평하다.
■ 화 (Shout Type)
가) Magnitude, Intensity가 가장 크다
나) 한 문장에 한 개의 Accented Point가 존재한다.
다) 어미 부분에서 음을 올리는 경우 1F가 높다(즉, 가벼운 느낌을 보임)
■ 웃음(Broad Laugh)
가) 소리가 깨끗하다 (Formant의 분포가 Uniform함)
나) 굵음의 정도는 개인차가 있다.
다) 동일한 모음이 반복된다.
라) 피치가 점진적으로 감소한다
■ 놀람(비명)
가) 1700Hz - 2500Hz에 에너지가 가장 많이 분포한다. (소리의 명확도가 높음)
나) 피치가 매우 높다

2.2 DRNN을 이용한 감성인식

2.2.1 특징추출

감성인식 시뮬레이터의 입력 값으로는 피치의 패턴들을 사용한다. 즉, 4가지 감정(평서,화,웃음,놀람)에 대한 대표 패턴들을 DRNN (Dynamic Recurrent Neural Network) 구조를 이용하여 학습,인식한다. 피치를 추출하기 위해서 우선 Autocorrelation Approach using Center-Clipping Function을 사용한다.

$$A(k)=\sum_{m=-\infty}^{\infty} x(m)x(m+k) \qquad (1)$$

(1)식은 Autocorrelation 함수를 나타내고, Center Clipping function 은 Autocorrelation 함수에 데이터를 입력시키기 전에 필요없는 정보를 제거하는 역할을 수행한다.

$$y(n)=c[x[n]] \qquad (2)$$

식 (2)와 그림 1. 이 Center clipping function을 나타낸다. 이 함수는 음성신호가 일정한 레벨(CL)내에 있으면 그 신호를 무시하고, CL 보다 크면 원래 신호에서 CL 을 뺀다. 이는 음성신호 중에서 피치에 해당하는 성분은 크기가 크게 나타나는 특징을 이용해서 잔여성분을 제거하는 방법이다. 그리고, CL은 프레임 내의 가장 큰 음성신호 레벨의 64%를 기준으로 한다[1].

그림 2 는 center clipping function을 사용하였을 때의 효과를 나타내는 것으로써 (a)는 원래의 신호를 나타내고 (c)가 clipped signal을 보여준다. (b)와 (d)는 각각 원래의 신호와 clipped signal에 대해 autocorrelation을 한 결과를 나타내는 그림으로써 clipped signal의 경우에 주기가 더욱 분명하게 나타나는 것을 확인 할 수 있다.

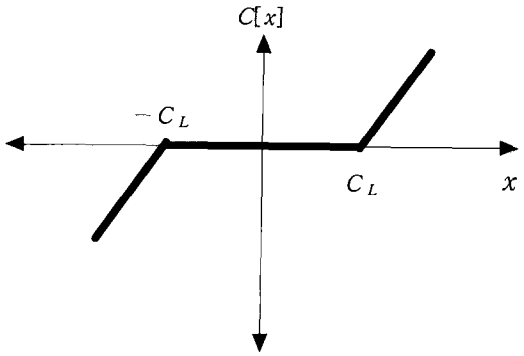

그림 1. Center clipping 함수
Fig. 1. Center clipping function

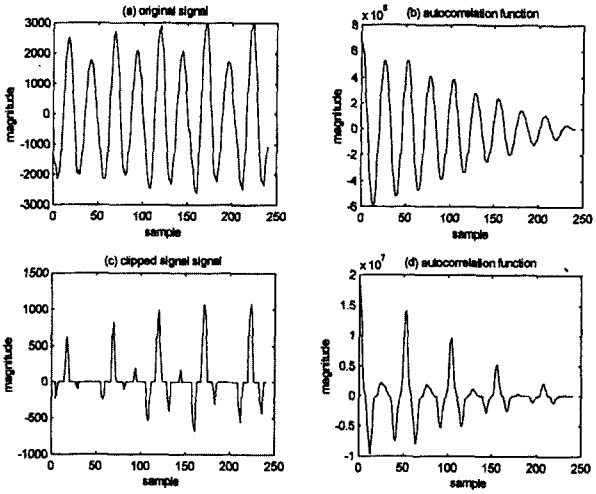

그림 2. Center clipping 함수의 효과
Fig. 2. The effect of applying the center clipping function

2.2.2 시뮬레이터

그림 3. 은 시뮬레이터의 구조를 나타낸다. 마이크를 통하여 음성이 입력되면, 음성의 피치를 추출하고 추출된 피치를 각 감정에 대응하여 학습을 시킨다. 학습을 통하여 신경망의 Weight를 획득 인식기를 통해 인식을 한다[2].

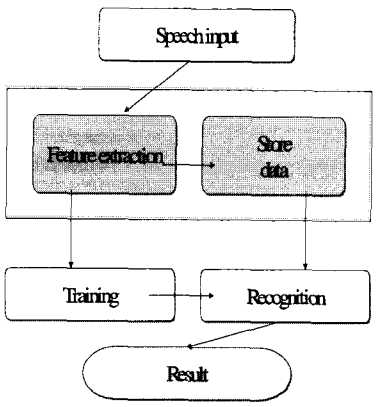

그림 3. 시뮬레이터의 구조
Fig. 3. Simulator structure

2.2.3 개체의 발생

개체를 발생시키는 알고리즘으로는 (1+100)-ES 를 사용하였다[3]. Evaluation function 은 목표 값과 결과값의 차를 사용하는 것과 Penalty Rule을 이용하는 방법 두가지를 이용하여 비교하였다.

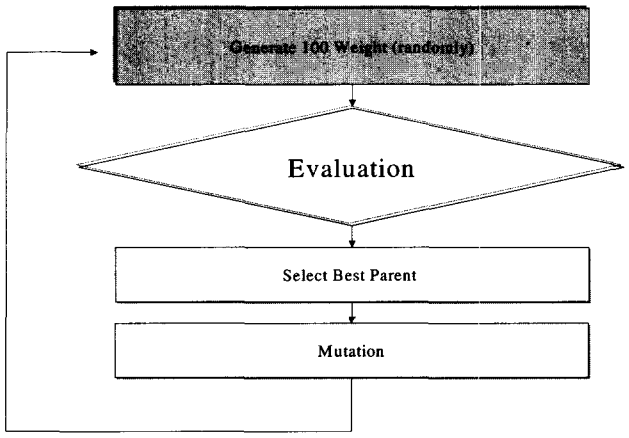

그림 4. (1+100)-ES
Fig. 4. (1+100)-ES

Penalty Rule은 목표 값과 결과 값의 차이의 정도에 따라 벌칙을 강화/약화 하는 방식이다. 목표 값과 결과 값은 4Bit 이고, 각각의 비트는 -1~1 사이의 실수 값을 사용하였다. 아래의 표에서 보는 바와 같이 -1을 가져야하는 비트가 0보다 크게 되는 경우 dif 값에 3을 곱해주는 벌칙을 적용하고 각 목표 값은 4비트 중 3비트는 -1이고 한 개의 비트만 1로 이뤄져있는데, 결과 값 중 그 비트들 간의 차이가 작아지면 그 것 또한 목표 값에서 멀어지는 것이므로 case 2 와같이 벌칙을 적용하였다.

표 5. Penalty 규칙의 적용
Table 5. Penalty rule

```
Case 1: if(goal1[sujja]==-1 && result1[sujja]>0)
            dif[0]*=3;

Case 2: For(sujja=0;sujja<=3;sujja++)
        if((result1[3]-result1[sujja])<0.3 && sujja!=3)
        dif=dif+3+(0.3-result1[3]+result1[sujja])*2
```

2.2.4 DRNN

DRNN은 뇌의 동적 특성을 모델링한 것으로써, 상호간 피드백 연결 되 있고, 지연 특성을 갖고 있다. 또한 교사학습 알고리즘을 사용하므로 가중치를 찾아내는 방법으로 보통 Back Propagation을 적용한다. 즉, 네트웍을 단순한 feed forward network으로 간주하고 표준적인 Back Propagation 알고리즘을 사용하는 것이다. Jordan's sequence generating network과 Elman's sequence prediction network이 자주 사용된다. 그렇지만, 이러한 방법들은 계산양이 많고 메모리를 많이 차지하는 단점이 있다. 본 논문에서는 계산이 상대적으로 단순하고 메모리 할당 양도 적은 (1+100)-ES를 사용하여 최적 개체를 찾았다[4].

- Fully connected.
- Input:1, Hidden:2, Out:4
- Input: Pitch.
- Out: Normal, Angry, laugh, Surprise

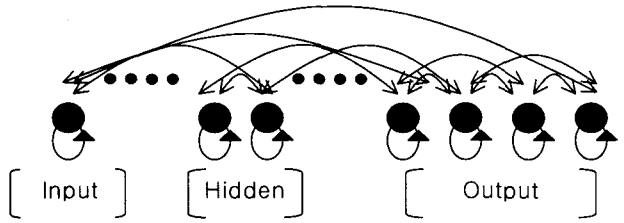

그림 5. DRNN의 구조
Fig. 5 Dynamic Recurrent Neural Network structure

Dynamic Recurrent Neural Network의 구조는 Fig 4. 와 같고, 음성이 입력이 시간에 따라 순차적으로 들어오므로 DRNN이 이러한 종류의 데이터에 적합하다.

i번째 뉴런의 출력은 다음 식과 같다.

$$y_i(t) = f(h_i(t-1)) + \Lambda(\sigma) \tag{3}$$

$$h_i(t) = (\sum_j w_{ij} y_j(t) + x_i(t)) \tag{4}$$

단, $h_i(t)$는 시간 (t-1) 에서 i 번째 노드에 대한 입력이다. $x_i(t)$는 시간 t에서의 외부 입력이다. $f(\cdot)$ 는 nonlinear derivative activation function 이다[5].

$$f(x) = \frac{2}{1+\exp(\frac{-2x}{u_0})} - 1 = \tanh(\frac{x}{u_0}) \tag{5}$$

2.3. 시뮬레이션 결과

2.3.1 입력의 종류에 따른 결과

시뮬레이터의 입력으로 피치와 포만트를 동시에 사용한 경우와 피치만 사용한 경우의 결과를 보면 표 6, 7과 같다. 포만트가 감성인식에서 의미를 갖기 위해선 문장의 내용에 따른 데이터베이스가 구축되어 있어야 하는데, 그러한 선행조건의 만족 없는 시뮬레이션은 오히려 결과에 악영향을 미친다.

표 6. 피치와 포만트를 입력으로 함께 사용한 경우
Table 6. When both pitch and formant are used as the input.

감정	인식결과 (성공회수/총시도수)
Normal	2/7
Angry	3/8
Laugh	2/4
Surprise	0/3

표 7. 피치만 사용한 경우
Table 7. Only pitch is used.

감정	인식결과 (성공회수/총시도수)
Normal	3/7
Angry	3/8
Laugh	3/4
Surprise	1/3

2.3.2 평가함수의 종류에 따른 결과

본 논문에서는 평가함수로써 목표 값과 결과 값의 단순 비교한 값을 이용하여 좋은 개체를 찾는 방법과 목표 값과 결과 값의 차가 클수록 Penalty를 주어 더욱 빨리 찾도록 하고, 하나의 결과 값 내에서도 0 과 1 두 비트가 적절한 자리에 있지 않은 경우 또한 Penalty를 주어 정확한 값에 빨리 가까워지도록 하는 Penalty rule을 적용하였다. 두 경우의 결과는 다음의 표 8,9에서 보여지는 바와 같다.

표 8. Raw difference로 평가한 경우
Table 8. Evaluated by raw difference

감정	인식결과 (성공회수/총시도수)
Normal	3/7
Angry	3/8
Laugh	3/4
Surprise	1/3

표 9 Penalty rule을 사용한 경우
Table 9. Evaluated by penalty rule

감정	인식결과 (성공회수/총시도수)
Normal	4/7
Angry	8/8
Laugh	3/4
Surprise	3/3

3. 결 론

본 논문은 음향적 요소들과 감성의 관계에 대해 분석하였고, 피치의 패턴을 입력으로 하여 시뮬레이터를 구성 결과를 살펴보았다. 표 9에서 보여진 결과는 꽤 고무적인 결과를 보여주는 편이다. 그러나, 이는 시행착오를 통한 파라미터의 조정이 필요했고, 어떤 학습 데이터를 사용하는지에 의존적이다. 즉, 피치만으로는 충분한 신뢰성을 갖고 있다고 할 수는 없다. 차후의 과제로는 본 논문에서 살펴보았던 음향적 요소의 분석결과를 이용하여 시뮬레이터에 보충 하는 것이다.

참 고 문 헌

[1] J. S. Han, Speech Signal Processing, Seoul, O-Sung media, p.90, 2000.

[2] C. H. Park, K. S. Heo, D. W. Lee, Y. H. Joo, and K. B. Sim, "Emotion Recognition based on Frequency Analysis of Speech Signal," *Proc. of the FIRA Robot 2002 Conference*, 2002.

[3] K. B. Sim, *Methodology of Artificial Life*, Seoul, Dream-Media, 2000.

[4] M. A. Arbib, *The Handbook of Brain Theory and Neural Networks*, The MIT Press, Cambridge, Massachusetts, pp. 796-799, 1995.

[5] H. B. Jun, D. W. Lee, D. J. Kim, and K. B. Sim, "Fuzzy Inference-based Reinforcement Learning of Dynamic Recurrent Neural Networks," *Proc. of SICE*, pp 1083-1088, 1997.

저 자 소 개

심귀보(Kwee-Bo Sim)

1984년 : 중앙대학교 전자공학과 공학사
1986년 : 동 대학원 전자공학과 공학석사
1990년 : The University of Tokyo 전자공학과 공학박사
2003년~현재 : 한국퍼지 및 지능시스템학회 부회장
2001년~2002년 : 대한전기학회 제어및시스템 부문회 편집위원 및 학술이사
2000년~현재 : 제어자동화시스템공학회 이사
1991년~현재 : 중앙대학교 전자전기공학부 교수

관심분야 : 인공생명, 진화연산, 지능로봇시스템, 뉴로-퍼지 및 소프트 컴퓨팅, 자율분산시스템, 로봇비전, 진화하드웨어, 인공면역계 등
Phone : +82-2-820 5319
Fax : +82-2-817 0553
E-mail : kbsim@cau.ac.kr

박창현(Chang-Hyun Park)

2001년 : 중앙대학교 전자전기공학부학사
2001년~현재 : 동 대학원 전자전기공학부 석사과정

관심분야: 진화연산, 신경회로망 등
Phone : +82-2-820-5319
E-mail : 3rr0r@alife.cau.ac.kr

주영훈(Young Hoon Joo)

1978년 : 연세대 전기공학과 졸업,
1984년 : 연세대 대학원 전기공학과 졸업.
1995년 : 동대학원 전기공학과 졸업(공박).
1986~1995년 8월 : 삼성전자(주) 자동화연구소(선임 연구원).
1998년 1월~1999년 2월 : 미국 University of Huston 전기 및 컴퓨터공학과 Post-doc.
2000년~현재 : 한국퍼지 및 지능시스템학회 편집이사
2001년~현재 : 대한 전기학회 제어계측분과 편집위원
1995년 9월~현재 : 군산대 공대 전자정보공학부 부교수

관심분야 : 퍼지제어, 지능제어, 유전알고리즘, 지능형 로봇
Phone : 063-469-4706
Fax : 063-469-4706
E-mail : yhjoo@kunsan.ac.kr