음성신호를 이용한 감성인식에서의 패턴인식 방법

The Pattern Recognition Methods for Emotion Recognition with Speech Signal

박 창 현, 심 귀 보*
(Chang-Hyun Park and Kwee-Bo Sim)

Abstract : In this paper, we apply several pattern recognition algorithms to emotion recognition system with speech signal and compare the results. Firstly, we need emotional speech databases. Also, speech features for emotion recognition is determined on the database analysis step. Secondly, recognition algorithms are applied to these speech features. The algorithms we try are artificial neural network, Bayesian learning, Principal Component Analysis, LBG algorithm. Thereafter, the performance gap of these methods is presented on the experiment result section. Truly, emotion recognition technique is not mature. That is, the emotion feature selection, relevant classification method selection, all these problems are disputable. So, we wish this paper to be a reference for the disputes.

Keywords : emotion recognition, ANN, bayesian learning, principal component analysis, LBG algorithm

I. 서론

사용자들은 그 시스템, 가상환경 혹은 로봇들에 사회적이면서 감정적인 반응을 보인다는 사실을 많은 연구자들이 발표했다. 예를 들어, 감성은 사용자가 가상 환경(Ijsselsteijn 2002;Lombard and Ditton 1997)을 인지할 때 영향을 미치고 그에 따라 반응한다(Dillion et al. 2000;Kalawsky 2000). 과학기술이 발전함에 따라 기계들의 중요성이 커졌다. 특히, 산업에서만 사용되었던 로봇들이 집에서도 사용되게 되었다. 그렇기 때문에 우리가 초점을 맞춰야하는 점이 있다. 즉, 인간은 매우 감정적이고 감정이 인간 사이의 상호관계에 중요한 역할을 한다는 점이다. 그래서 로봇들 또한 인간의 감정이 섞인 명령을 인식할 수 있도록 개발 되어야한다. 많은 연구자들이 이 분야에 대해 연구해왔다. Fatma et al (2003)은 Multimodal affective user interface를 구현했고 감성과 관련된 생체신호를 분석했다[1]. V.Hozjan et al(2003)은 음성에 대해 내용독립적인 감성 인식을 시도했는데, 특히 주목할 점은 여러 언어에 대해서 실험했다는 점이다. 이 연구는 감성 분류를 위해 인공 신경망을 이용했다. 1998년에 Chen 과 Tao 는 6개의 감성(행복, 슬픔, 위험, 싫어함, 놀람 그리고 두려움)들을 피치와 각 문장의 RMS 에너지 포락선을 이용하여 구분했다. 하지만 이 특징들은 감성을 인식하는데 적절한 성능을 보이지 못했기 때문에 표정인식을 추가하여 성능을 높일 수 있음을 보였다. 정리하면, 오디오 정보와 이미지 정보를 함께 사용할 때 인식 성능이 더 좋아질 수 있다는 것이다. J. Nicholson은 감성을 의식적인 것과 무의식적인 것으로 나누었고 인식하기에 훨씬 쉽다는 이유로 의식적인 감성에 초점을 맞추어 연구를 진행하였다. 또한, 그는 8개의 감성(joy, tease, fear, sadness, disgust, anger, surprise and neutrality)들의 특징들을 추출했고 그 특징들은 운율적인 특징과 음성적인 특징들로 분류되었다. 특히, 그는 하위 신경망을 decision logic에 적용하여 결과를 얻었다. 본 논문은 음성을 사용한 감성 인식 시스템에 여러 가지 패턴 인식 알고리즘을 적용하고 그 결과를 비교할 것이다. 그렇게 하기 위해서는 먼저, 감성적인 음성 데이터베이스가 필요하다. 또한 감성 특징 점 추출이 데이터베이스 분석단계에서 결정 되어야 한다. 두 번째로 인식알고리즘이 이 음성 특징들에 적용된다. 우리가 시도하는 알고리즘들은 인공신경망, 베이지안 학습, 주성분분석법, LBG 알고리즘이다. 그리고 나서 이 방법들의 성능 차이가 실험 결과 부분에서 보여질 것이다. 본 논문의 구성은 다음과 같다. 먼저 2절에서는 feature extraction을 다룬다. 추출된 feature는 acoustic feature로써 언어적 특성은 배재하였다. 3절에서는 본 논문에서 사용된 패턴 인식 알고리즘인 신경망, 베이지안 학습, 주성분 분석법을 이용한 패턴 분류, LBG 알고리즘에 대한 설명과 본 연구에서의 설정에 대해 보여준다. 4절은 emotion speech database 구축에 대한 부분으로써 실제로 감정적 음성 샘플을 수집하기위한 실험 환경에 대한 설명이다. 5절은 위와 같은 준비를 바탕으로 각각의 패턴인식 알고리즘을 사용하여 감성 인식을 수행한 결과를 보여준다. 6절에서는 결론으로 마무리 짓는다.

* 책임저자(Corresponding Author)
논문접수 : 2005. 11. 15., 채택확정 : 2006. 2. 5.
박창현, 심귀보 : 중앙대학교 전자전기공학부
(3rror@wm.cau.ac.kr/kbsim@cau.ac.kr)
※ 본 연구는 산업자원부의 뇌정보처리에 기반한 감각정보 융합 및 인간행위 모델 개발사업의 연구비 지원으로 수행되었음. 연구비 지원에 감사드립니다.

II. Feature Extraction

감성 인식기는 두 부분으로 구성되어 있다. 첫 번째 부분은 음성으로 특징을 추출하는 부분이고 두 번째 부분은 그 특징들을 이용하는 패턴 인식 부분이다. 특징들은 피치의 통계치, 소리의 크기, 섹션개수 등이다. 피치 추출 방법으로는 가장 일반적인 방법들 중 하나인 autocorrelation approach를 사용했다. 피치 값은 0.1초마다 추출 되어졌고

그 값들의 평균이 pitch mean으로 정의 했다. 그리고 분산 값 또한 동일한 데이터에서 얻어졌다. 소리의 크기는 magnitude estimation method에 의해서 구해졌고, 섹션 개수, Increasing Rate(IR), Crossing Rate(CR) 등은 우리의 이전 논문에서 사용한 방법으로 구했다[8].

III. Emotion Recognition에 사용할 패턴인식 알고리즘

1. Artificial Neural Network(ANN)

ANN은 실수, 이산 값이나 벡터를 학습 하는데 있어서 매우 일반적이고 실용적인 방법이다. Backpropagation(BP) 같은 알고리즘들은 입출력 학습 쌍들을 최적으로 하기 위해 네트웍 파라메터들을 조정하는데 이때 gradient descent 알고리즘을 사용하는 것이다. ANN 학습은 학습 데이터의 에러들에 강인하고 이미지 패턴 분류, 음성인식 그리고 로봇제어 전략 학습 같은 문제들에도 잘 적용되어 왔다. 이 논문에서 사용된 ANN의 파라메터들은 다음과 같다.

우리가 다루려고 하는 문제가 2개의 이진 출력들을 갖고 있기 때문에 25%이하의 error tolerance 가 올바른 출력을 결정하는데 적절하다.

2. Bayesian Learning

베이즈의 이론은 사전확률에 근거해 가설의 확률을 계산하는 방법을 제공한다. 그렇기 때문에 해당 문제에 대한 사전확률이 구해져야 한다. 사전확률을 위해서 400개의 샘플들이 사용 되어졌다. 좀 더 상세히 기술하면, 400개의 샘플들에 대해서 각 감정과 특징 들(피치 평균, 소리의 크기, 섹션 개수 등)간의 관계를 관찰한 것이다. 다음의 그래프들은 그 결과를 보여준다.

그림 1은 섹션 개수에 대해서 각 감정의 분포 확률을 나타낸다. 그림에서 보여 지는 것처럼 depress의 경우는 다른 감정들과 확연히 분류되는 것을 알 수 있다. 그리고, normal, happy, angry 각각은 서로 약간씩 다른 영역을 갖고 있다는 것 또한 확인 할 수 있다. 그림 2는 피치 평균에 대한 각 감정의 분포 확률을 나타낸다. 이 그림에서도 각 감정들이 고유의 영역을 갖고 있음을 알 수 있다. 그림 3은 소리 크기에 대한 감정별 분포 확률을 나타내는 것인데, 이 경우 depress의 경우 분명히 다른 영역을 갖고 있지만 나머지 감정들의 분포 영역의 구분이 모호하기 때문에 이 분포만으로는 분류가 어렵다. 하지만 3가지 분포확률을 혼합하면 좋은 결과를 얻을 수 있을 것으로 기대된다.

3. Principal Component Analysis(PCA)

PCA는 유명한 특징 추출 방법이고 패턴 인식 기술로도 많이 사용되어 왔다. 특히, 얼굴 인식 같은 많은 컴퓨터 비전 응용 프로그램들에서 사용되어왔다. 그림 4는 감성 데이타를 학습하는 과정을 보여준다. 감성 특징 벡터는 피치 평균, 소리 크기, 섹션 개수, IR, CR로 구성 되었다. 즉, 5×1 벡터가 수집되고 학습 벡터 집합인 S 가 구성된다. 그리고 나서, 학습이 그림 4의 과정에 따라서 수행되었다. 그리고 새로운 입력 데이터를 분류하기위한 과정이 다음과 같다. 첫째, 다음과 같이 고유 데이터 요소들을 구한다. 둘째로, 고유 데이터 벡터를 구한다. 마지막으로, 학습 데이터와 입력 데이터간의 거리를 측정한다.

4. LBG 알고리즘

유클리디안 거리를 사용하면 간단히 클러스터들의 선형 경계들은 알 수 있다. 즉, K-means를 통하여 클러스터들을 찾고 어떤 두 클러스터의 중심을 연결하여 이 직선을 수직 이등분하는 선분이 최적 경계가 된다. 그리고 각 클러스터의 중심에서 가장 가까운 모든 점들을 해당 클러스터에 포함시키면 된다. 분류를 하는 알고리즘이다. K-means는 초기 중심들의 선택에 민감한 특징을 가진다. 그러므로 K-means

표 1. 신경망 파라메터 설정.

Table 1. Parameter setting of neural network.

Parameter	Values
Input Units	3~5
Hidden Units	11
Output Units	2
Learning Rate	0.003
Tolerance	0.25
Sigmoid Function	$\frac{1}{1+e^{-3x}}$

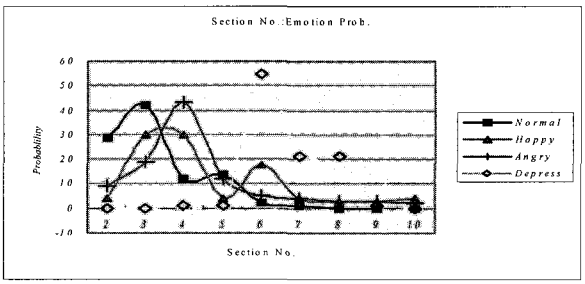

그림 1. Sect.No에 따른 감정별 빈도수.

Fig. 1. PDF of Sect.No for each emotion.

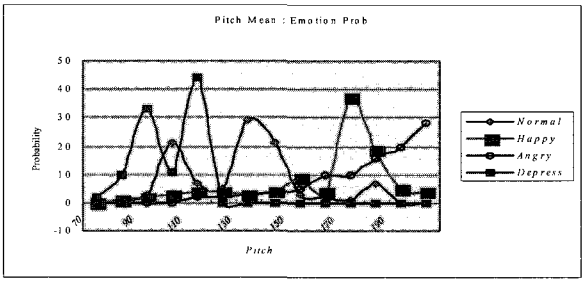

그림 2. 피치평균에 따른 감정별 빈도 수.

Fig. 2. PDF of pitch mean for each emotion.

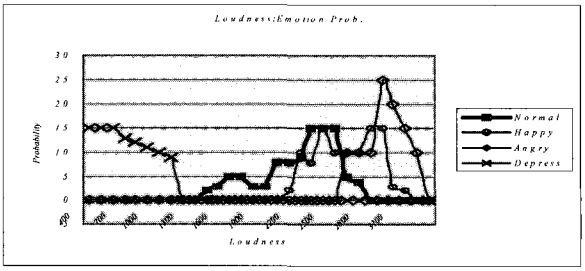

그림 3. Loudness에 따른 감정별 빈도 수.

Fig. 3. PDF of loudness for each emotion.

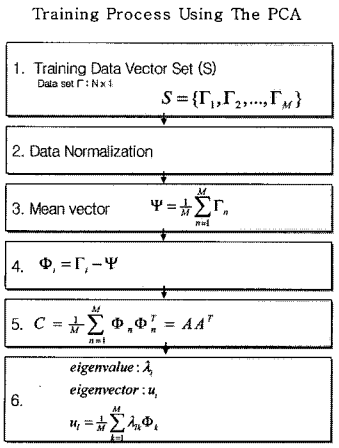

그림 4. PCA를 이용한 학습과정.

Fig. 4. PCA algorithm.

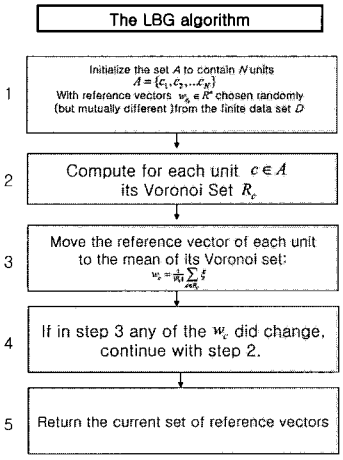

그림 5. LBG 알고리즘.

Fig. 5. LBG algorithm.

의 초기값을 랜덤 데이터 점으로 하는 대신에 이진 분리로 구한 중심을 사용하면 표준 K-means 방법을 사용하는 경우보다 더 나아질 것이다. 이러한 알고리즘을 LBG 알고리즘이라고 한다.

IV. Emotion Speech Database 구축

10명의 남성 대학원생들(나이:24~31)에게 4가지 감성으로 총 400개의 음성 샘플을 얻었다. 그들은 보통의 한국 사람이고 여러 지역의 출신으로 이루어져 있다. 녹음된 형태는 11KHz, 16bit, mono 이고 소리의 크기가 피험자와 마이크의 거리에 따라 달라질 수 있으므로 그 거리를 10Cm 로 고정하였다. 녹음된 문장들은 30개의 일상적이고 단순한 것들이었고 문장의 길이는 6~10음절로 제한해 놓았다. 30 개의 미리 준비된 문장들은 그것들을 감성 데이터로 채택해도 될지 확인을 받아야 하기 때문에 녹음한 사람들 이외의 다른 30명에게 "녹음된 소리가 어떤 감정을 포함하고 있는 것 같은가?" 라는 질문을 해서 90%의 동의를 얻은 10개의 문장에 대해서 녹음을 하였다.

V. Experimental Result

1. Recognition with Artificial Neural Network(ANN)

그림 6은 앞의 설정대로 하였을 때 31000번 반복한 에러 파형이다. 또한 table 2는 학습 완료시점에서의 출력 노드에서 관찰된 결과를 나타낸다. 학습은 에러가 0.009167인 점에서 종료되었다. 그림 7은 400개의 speech samples에 대해서 학습 결과를 적용한 결과이다. 각 특징의 유용함을 확인해보기 위해서 4가지의 특징 집합들이 ANN에 입력되어 테스트 되었다. 그 집합들 중에서 feature2(pitch mean, loudness, CR)이 가장 잘 인식된 집합으로 판명되었다.

2. Recognition with Bayesian Learning

표 3는 400 samples에 대해서 테스트한 결과이고 사전확률은 앞 절에서 얻은 확률분포를 사용하였다. 표 3를 보면 인식율이 사람마다 그리고 감정별로 다른 것을 확인할 수 있다. 그 이유는 사람마다 감성을 표현 하는 방식이 조금씩 다르기 때문이다. 하지만, 비선형적인 데이터 분포에 흡사하게 분류가 된 알고리즘의 특성 때문에 비교적 높은 인식율을 보였다.

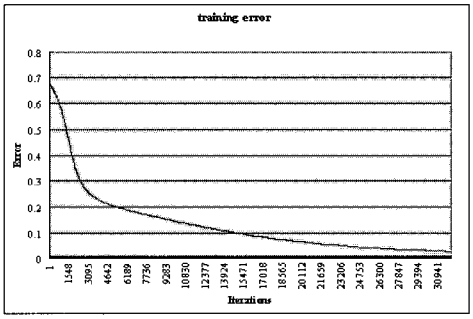

그림 6. 학습 에러 그래프.

Fig. 6. Error graph.

표 2. 신경망 학습 결과.

Table 2. Training result.

Emotion	Expression Pattern
(0)0.001813 (0)0.017775	(0)0.063257 (1)0.917163
(1)0.981922 (1)0.999878	(0)0.069832 (1)0.955858

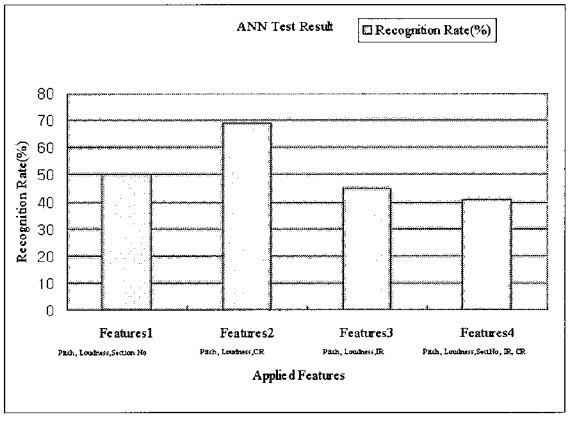

그림 7. ANN 테스트 결과.

Fig. 7. ANN test result.

3. Recognition with PCA

PCA 알고리즘에 선택된 입력 파라메터는 Loudness와 Pitch mean이다. 모든 학습 샘플들이 PCA에 입력되지는 않고 대신 4개의 감정들 각각의 대표 샘플들만이 LBG 알고리즘에의해 선택되었다. 그리고, 100개의 테스트 데이터를 임의로 선택하였고 실험이 3번 수행되었다. 실험의 결과는 LBG의 결과와 비슷하게 나왔다. 표 4는 3번의 실험에 대한 결과를 보여준다.

표 3. BL을 이용한 감성 인식 결과 : 인식율.
Table 3. Recognition rate of BL.

	Normal	Happy	Angry	Depress	Average
S1	57%	40%	80%	70%	62%
S2	90%	73%	80%	94%	84%
S3	70%	51%	89%	91%	75%
S4	67%	56%	91%	85%	75%
Average	71%	55%	85%	85%	74%

표 4. PCA를 이용한 감성 인식결과 : 에러율.
Table 4. Error rate of PCA test.

	Normal	Angry	Depress	Happy	Average
1st trial	68%	22%	18%	44%	38%
2nd trial	73%	18%	15%	50%	39%
3rd trial	70%	21%	20%	40%	37.75%

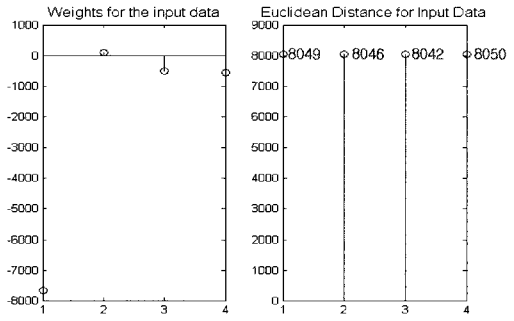

그림 8. PCA를 이용한 인식 결과: 3번째 데이타와 비슷한 패턴을 입력한 결과로써 3번째 막대의 길이가 가장 짧으므로 올바른 인식 결과를 보여준다.
Fig. 8. Recognition result graph of PCA.

표 5. LBG 알고리즘을 사용한 인식결과 : 에러율.
Table 5. Error rate of LBG test.

	Normal	Angry	Depress	Happy	Average
Loudness, Pitch Mean	71.4%	22%	20%	40%	38.35%
Sect.No./Syllables no, Pitch mean	57%	71%	20%	60%	52%
Sect.No./Syllables no, Loudness	71%	22%	20%	40%	38.25%

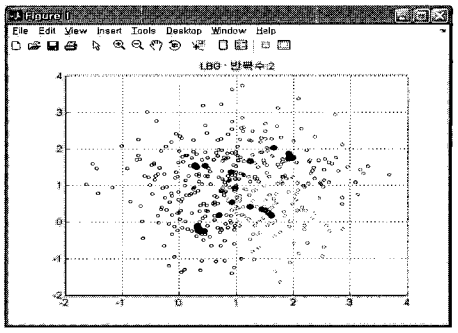

그림 9. LBG 알고리즘에의해 최종적으로 분류된 데이터.
Fig. 9. LBG test result graph.

4. Recognition with LBG

본 논문에서 LBG 알고리즘은 3쌍의 학습 파라메터 타입을 채택해서 테스트되어졌다. 3쌍의 조합은 (Loudness, Pitch mean),(Sect.no./Syllables no., Pitch mean) 그리고 (Sect.no./Syllables no., Loudness). 여기서, 는 발화의 템포를 의미하는 것이다. 이러한 3쌍의 파라메터를 채택하여 실험한 결과는 표 5와 같다.

6. Conclusion

본 논문은 4가지 알고리즘을 사용하여 감성 인식에 적용한 결과를 비교하였다. 감성 인식을 위한 매체로는 음성, 표정, 생체 신호등이 있으나 본 논문에서는 음성 신호로부터 감성 특징점을 추출하여 4가지 알고리즘에 적용, 테스트하였다. 사용된 특징점은 신경망과 LBG 알고리즘을 사용하여 인식율이 좋은 집합으로 선택된 것이고 그 결과 베이지안 학습과 신경망을 통한 인식 결과가 PCA나 LBG를 사용한 결과보다 좋았다.

참고문헌

[1] F. Nasoz, K. Alvarez, C. L. Lisetti and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," Springer-verlag, london, 2003.

[2] V. Hzjan and Z. Kacic, "Context-independent multilingual emotion recognition from speech signals," *International Journal of Speech technology*, pp. 311-320, 2003.

[3] L. S. Chen, H. Tao, T. S. Huang, T. Miyasato and R. Nakatsu, "Emotion recognition from audiovisual information," *IEEE Second Workshop on Multimedia Signal Processing*, 1998.

[4] J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion recognition in speech using neural networks," *Proc. of ICONIP*, vol. 2, 1996.

[5] S. Batliner, K. Fisher, R. Hyber, J. Spilker and E. Noth, "Desperately seeking emotions : actors, wizards and human beings," *Proceedings of the ISCA Workshop on Speech and Emotion.*

[6] T. Moriyama and S. Ozawa, "Emotion recognition and

synthesis system on speech," *IEEE International Conference on Multimedia Computing and Systems*, vol. 1, 1999.

[7] D. Galanis, V. Darsinos and G. Kokkinakis, "Investigating emotional speech parameters for speech synthesis," *Proc. of ICECS*, vol. 2, pp. 13-16, Oct, 1996.

[8] C. H. Park, K. S. Byun and K. B. Sim, "The implementation of the emotion recognition from speech and facial expression system," *Proc. of ICNC*, Part 2, pp. 85-88, Aug, 2005.

[9] T. M. Mitchell, *Machine Learning*, McGraw-Hill International Edition ,Singapore, 1997.

[10] R. Rojas, *Neural Networks a Systematic Introduction*, Springer, Germany, 1996.

[11] J. Rogers, *Object-Oriented Neural Networks in C++*, Academic Press, USA, 1997.

[12] R. O. Duda, P. E. Hart and D. G. Stork, Pattern Classification, A Wiley Interscience Publication, USA, 2001.

박 창 현

2001년 중앙대학교 전자전기공학부 졸업. 2003년 동 대학원 전자전기공학부 석사. 2003년~현재 동 대학원 박사과정 재학 중. 관심분야는 패턴인식, 기계학습, 진화연산 등.

심 귀 보

1956년 9월 20일생. 1984년 중앙대학교 전자공학과(공학사). 1986년 중앙대학교 전자공학과(공학석사). 1990년 동경대학교 전자공학과(공학박사). 1991년~현재 중앙대학교 전자전기공학부 교수. 2003년~2004년 일본 계측 자동제어학회(SICE) 이사. 2000년~2004년 제어 · 자동화 · 시스템공학회 이사. 2002년~현재 중앙대학교 산학연컨소시엄센터 센터장 및 기술이전센터 소장. 2005년~현재 한국퍼지 및 지능시스템학회 수석부회장. 2006년~현재 한국퍼지 및 지능시스템학회 회장. 2005년 제어 · 자동화 · 시스템공학회 Fellow 회원. 관심분야는 인공생명, 지능로봇, 지능시스템, 다개체 시스템, 학습 및 적응알고리즘, 소프트 컴퓨팅(신경망, 퍼지, 진화연산), 인공면역시스템, 침입탐지시스템, 진화하드웨어, 인공두뇌, 지능형 홈 및 홈네트워킹, 유비쿼터스 컴퓨팅 등.