화자 의존 환경의 AMR 7.4Kbit/s모드에 기반한 보코더

정회원 민 병 제*, 박 동 철**

A New Vocoder based on AMR 7.4Kbit/s Mode for Speaker Dependent System

Byung-Jae Min*, Dong-Chul Park** ***Regular Members***

요 약

본 논문은 AMR(Adaptive Multi Rate)코더의 7.4kit/s 모드를 기반으로 화자 의존적인 환경에서 더욱 압축률을 높인 새로운 켈프(CELP)계열의 코더를 제안한다. 제안된 코더는 OGM(OutGoing Message)이나 TTS(Text-To-Speech) 등 한 사람의 음성만을 필요로 하는 시스템에서 유용하게 사용할 수 있다. 새로운 코더의 압축률을 높이기 위해서 무감독 학습 신경망인 Centroid Neural Networks(CNN)를 이용한 새로운 LSP 코드북을 생성하여 사용한다. 또한 고정 코드북 탐색 단계에서 AMR 7.4 kbit/s 모드에서는 4개의 펄스를 서브프레임 마다 사용하는 대신에 새로운 코더에서는 오직 2개의 펄스만을 사용하기 때문에 압축률을 더 높일 수 있다. 이로 인해서 스피치의 질이 감소하게 되는데, 각 서브프레임 마다 예상하는 펄스를 적용함으로써 보상 받을 수 있다. 제안된 보코더는 기존 AMR 7.4Kbps모드와 비교해 27% 높은 압축률을 가지는 동시에, MOS(Mean Opinion Score)의 면에서 볼 때, 대등한 음질을 보였다.

Key Words : Linear Predictive Coding, Centroid Neural Network, Adaptive Multi Rate

ABSTRACT

A new vocoder of Code Excited Linear Predictive (CELP) based on Adaptive Multi Rate (AMR) 7.4kbit/s mode is proposed in this paper. The proposed vocoder achieves a better compression rate in an environment of Speaker Dependent Coding System (SDSC) and is efficiently used for systems, such as OGM(Outgoing message) and TTS(Text To Speech), which needs only one person's speech. In order to enhance the compression rate of a coder, a new Line Spectral Pairs(LSP) code-book is employed by using Centroid Neural Network (CNN) algorithm. In comparison with original(traditional) AMR 7.4 Kbit/s coder, the new coder shows 27% higher compression rate while preserving synthesized speech quality in terms of Mean Opinion Score(MOS).

Ⅰ. 서 론

음성의 압축은 효과적인 전송과 저장 공간에 대한 방법의 최적화를 위해 간결한 디지털 표현을 얻는 것을 목적으로 한다. 실제적인 음성 압축의 목적은 전송과 저장시스템, 암호화, 복호화의 과정을 거쳐 얻게 되는 합성된 음성 신호의 질의 감소가 없이 비트 레이트를 감소시키는 것이다[1]. 보통 더 높게 음성을 압축하게 되면 음성의 왜곡은 증가하고 음성의 질이 떨어지게 된다. 음성 신호를 압축하기 위해서

※ 본 연구는 한국과학재단 특정기초연구(R01-2007-000-20330-0)지원으로 수행되었음.

* (주) SFA Engineering, ** 명지대학교 정보공학과 지능컴퓨팅 연구실

논문번호 : KICS2008-02-084, 접수일자 : 2008년 2월 12일, 최종논문접수일자 : 2008년 9월 3일

많은 음성 코딩 연구가 진행 되었고, 그 결과 band width가 제한된 어플리케이션을 위한 적합한 음성 압축의 많은 기술들이 발표되었다[2-8].

현대의 중요한 음성 압축 기술 모델 중 하나는 Code Excited Linear Predictive(CELP) 모델이다. CELP 코딩의 프레임웍은 [9]에서 제안되었고 낮은 비트 레이트에서 높은 질의 음성을 만들어 낼 수 있다[10]. 그러나 초기의 CELP모델은 계산의 복잡성으로 인해 실시간에서 구현되기엔 시간상으로 지연(delay)이 문제가 되었다. 이러한 CELP모델을 기본으로 계속 많은 연구가 진행됨으로 해서 CELP계열 코더 모델에서 최적화된 AMR Vocoder가 개발되었다.

최근 핸드폰의 발달로 인해 사용자 안내 멘트(OutGoing Mssage)나 텍스트를 읽어주는 TTS(Text-To-Speech)[11-12] 기능이 많이 탑재된다. 그러나 핸드폰의 소형화에 따른 저장 장치의 한계로 인해 현대의 핸드폰 시스템은 음성의 질이 왜곡되거나 감소되지 않는 보다 높은 압축률의 코더를 필요로 한다. 본 논문은 이러한 화자 의존적인 시스템[11]에서 기존 상용화 코더인 AMR 7.4Kbit/s모드를 기반하여 보다 높은 압축률의 보코더의 개발을 목적으로 한다.

본 논문의 Ⅱ장에서는 AMR 보코더에 대해서 알아보고 Ⅲ장에서는 CNN알고리즘에 대하여 검토해 본다. Ⅳ장에서는 음성 신호의 질을 높이기 위한 고정 코드북 검색 단계에서의 예상하는 펄스에 대해 알아보고 Ⅴ장에서는 새로운 보코더의 모델링 방법에 대하여 설명한다. Ⅵ장의 실험과 결과를 통해 새로운 코더의 우수성을 입증하고 Ⅶ장에서는 결론을 내린다.

Ⅱ. AMR(Adaptive Multi Rate) 코더

ETSI와 3GPP에서 차세대 이동통신 IMT-2000 서비스의 음성부호화기의 표준으로 채택한 AMR 음성 코덱은 음성을 부호화 하여 전송하는데 필요한 전송률이 4.75kbit/s로부터 12.2kbit/s인 8개의 음성 부호화기로 합쳐놓은 형태이다[13]. 그리고 여기 신호의 모델링을 위해Algebraic Code Excited Linear Prediction 방식을 이용한 Multi Rate ACELP를 사용한다. 이 부호화기는 매 20ms마다 전송률의 조절이 가능하며, 네트워크적 입장에서 볼 때 적절한 휴대 전화 통화 품질을 보장하면서 최대의 사용자를 서비스 하는 방안으로 이용될 수 있다. 각 160 샘플들에 대해 음성 신호는 CELP모델[14]의 파라미터(LP filter coefficients, adaptive and fixed codebook's indices and gains)를 추출하기 위해 분석된다. LP분석은 프레임당 12.2Kbps는 모드에 대해 두 번 나머지 모드에 대해서는 한번만 수행한다. 12.2kbps 모드에 대해, LP 파라메터 셋이 LSP로 변환되고 split vector quantization(SVQ)를 사용하여 양자화 하다. 하나의 음성 프레임은 각 40 샘플, 5ms인 4개의 서브 프레임으로 나누어진다. 적응/고정 코드북 파라메터는 매 서브 프레임 마다 전송된다. 양자화 된/양자화 되지 않은 LP 파라메터 또는 그들의 보간된(Interpolated)버전들은 서브 프레임에 의존하여 사용된다. 개루프 피치 지연(Open-Loop Pitch Lag)은 그 밖의 모든 부 프레임 마다 예측된다.

피치 지연의 값을 결정은 지각 가중된(Perceptually Weighted) 음성 신호 영역에서 이루어진다. 개회 피치 분석으로 개회로 피치 지연 값이 구해지면 그 값의 부근에서 폐회로 피치 분석(Closed-Loop)을 통해서 피치 지연(delay) 값과 그 이득(gain)이 계산된다. 이때 피치 지연 값은 보간(Interpolation)을 통해서 1/6 혹은 1/3의 정확도 까지 계산된다. 이 단계가 지나면 마지막으로 고정 코드북 검색을 한다. 고정 코드북은 Interleaved Single-Pulse Permutation(ISPP)구조에 기반한다. 고정 코드북 검색 과정을 통해 최적의 여기신호를 구성 할 수 있는 펄스의 위치와 부호가 결정되고 그에 따른 이득 값을 얻을 수 있다[14,15].

Ⅲ. CNN (Centroid Neural Network) Algorithm

CNN 알고리즘은 기존 k-means 알고리즘을 기초로, 주어진 데이터에 존재하는 군집의 중심을 찾는다. 기존 CNN은 승자, 패자의 연결강도의. 갱신 방법에 의해 지역적으로 최적의 연결 강도를 설정함으로 주어진 데이터를 표현할 수 있는 중심을 설정하는데 탁월하며, 또한 사전에 학습계수나 전체 학습 Epoch수를 설정할 필요가 없다. 입력 벡터 $\vec{x}$가 n 시간에 인가된 경우, 승자 뉴런 j와 패자뉴런 i의 연결강도 갱신은 아래와 같다.

$$w_i(n+1) = w_i(n) + \frac{1}{N_i}[x(n) - w_i(n)]$$

$$w_j(n+1) = w_j(n) + \frac{1}{N_j}[x(n) - w_j(n)]$$

위 식에서 w_i와 w_j는 각각 승자뉴런과 패자뉴런의 연결강도를 표현하며, N_i와 N_j는 각 군집 i와 j의 데이터 수를 나타낸다. 출력 뉴런의 연결강도는 아래의 식과 같이 총 거리를 최소화 하는 방법을 선택한다.

$$w_j = \min_{w} \sum_{i=1}^{N_i} \| x_j(i) - w \|^2$$

위 식에서 N_i는 i 군집에 속한 데이터 수를 나타낸다.

CNN에 관한 더 자세한 설명은 [16-18]에서 찾을 수 있고 CNN 알고리즘을 SOM[19] 알고리즘을 포함한 기존의 무감독 학습 신경망 알고리즘과 비교하면, CNN 알고리즘이 더 좋은 결과를 보이며, 더 적은 계산량을 요구하고 다른 알고리즘보다 더 안정적으로 수렴한다.

Ⅳ. 이전의 고정 코드북을 이용하는 예상하는 펄스

압축률이 높은 코더의 디자인을 위해서는 한 프레임 당 전송하는 bit의 양을 상대적으로 감소시켜야 한다. 한 프레임을 표현하기 위해서 사용되는 bit의 양에서 가장 많은 비중을 차지하는 것이 고정 코드북 검색 단계를 통해 얻어진 고정 코드북이다. 이러한 고정 코드북 검색은 서브 프레임 별로 이루어지는데 AMR 4.75Kbit/s 모드에서는 단 두 개의 펄스만을 이용하여 하나의 서브프레임을 신호를 표현하게 된다. 이 고정 코드북의 펄스가 많아질수록 스피치의 질이 좋아지는 반면 전송하게 되는 bit의 양은 증가하게 된다. 적은 펄스를 사용함으로써 음성의 질이 떨어지는 단점을 보완하기 위해 아래의 설명과 같이 전송해야하는 bit의 양은 증가하지 않으면서 서브프레임마다 펄스의 수를 더 많이 사용할 수 있게 서브프레임마다 예상하는 펄스를 추가한다[20].

전체의 여기 신호를 표현하기 위해서 적응 코드북을 구하는 검색 과정이 끝나면 그 다음으로 고정 코드북 파라메터를 얻기 위한 고정 코드북 검색 과정을 실행한다. 고정 코드북 검색 단계에서 목적(target) 신호는 적응 코드북 검색 시 목적 신호에서 적응 코드북 검색 후 만들어진 적응 코드북으로 합성한 신호와의 차이 값으로 한다. 이 고정 코드북 검색 과정의 목적 신호를 살펴보면 적응 코드북 검색 단계의 목적 신호와 유사한 모양의 성질을 발견 할 수 있다. 그래서 적응 코드북 검색 과정에서 사용한 피치 지연만큼의 이전 고정 코드북의 펄스의 위치를 현재 서브프레임의 고정 코드북에 추가함으로써 신호의 질을 높일 수 있다. 각 서브 프레임에서 전체 고정 코드북 C_s는 다음과 같다.

$$C_s = g_f \cdot (C_p + C_f)$$

C_p는 예상하는 펄스의 위치이고 C_f는 기존 AMR 고정 코드북 검색 단계에서 만들어진 고정 코드북 벡터이다. g_f는 고정 코드북 전체의 이득을 나타낸다. 그림 1은 예상하는 펄스를 계산하는 과정을 도식화 한 그림이다.

고정 코드북 검색 단계에서 전체 여기신호C_s는 k의 길이만큼의 버퍼 b에 저장이 된다. 이것은 적응 코드북 검색 과정에서 적응 코드북 버퍼에 과거의 전체 여기 신호를 저장하는 것과 유사하다. 현재의 서브 프레임을 위한 예상하는 펄스 C_p의 선택은 길이 N의 움직이는 윈도우에 의해서 검색 된다. 현재 서브 프레임의 피치 지연 측정값에 의해서 결정되는 위치에 해당하는 버퍼 b에서 예상하는 펄스를 현재 서브 프레임에 추가함으로써 고정 코드북을 더 많은 펄스로 표현 할 수 있게 한다. 예상하는 펄스는 다음과 같이 표현된다.

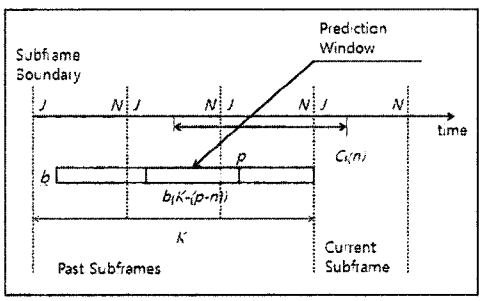


그림 1. 예상하는 펄스의 계산 과정[15]

$$C_p(n) = \begin{cases} b(K-(p-n) & n \le p \text{ and } n \le N \\ 0 & n > p \text{ and } n \le N \end{cases}$$

Ⅴ. 제안하는 새로운 보코더

제안하는 새로운 모델은 화자 의존 환경에서 AMR 7.4kbit/s모드를 기반으로 디자인 된다. 화자 의존 환경의 특성은 한 사람의 화자로부터 얻어진 입력 음성 신호만을 사용한다는 것이다. 그러므로 입력 음성 신호들은 개개인적인 특성을 가지게 되고, 그 음성들의 특성은 주파수 적으로 유사한 포락선을 보이게 된다. 그 결과로서, 음성 신호를 해석할 때 LP 분석을 통해 얻어진 LSP벡터는 주파수적인 정보가 비슷하게 나타난다. 제안된 모델에서는 LSP 벡터를 전송하는 과정에서 AMR에서 사용하는 SMQ(Split matrix Quantization) 방법을 사용하는 대신에, 코더의 생성에 많은 이점을 지니는 CNN알고리즘을 통한

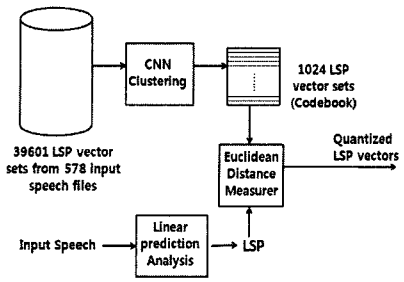


그림 2. 벡터 양자화 된 LSP 전송 과정의 블록 다이어그램

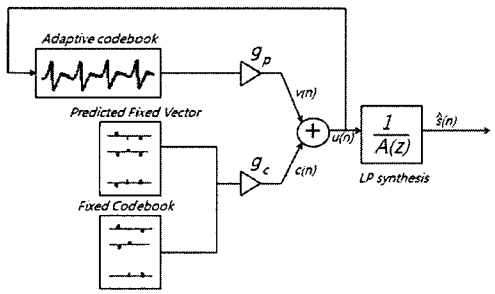


그림 3. 새로운 보코더의 간단한 블록 다이어그램

군집화된 개체들을 이용해서 양자화 된 코드북을 생성한다 [21].

또한, 현재 프레임에서 생성된 LSP 벡터와 거리 계산에 의한 가장 가까이에 위치한 LSP 코드북 인덱스를 전송한다. 이러한 양자화의 사용은 전송되는 LSP 벡터들이 작은 코드북에 의해 나타나기 때문에 더 높은 압축률을 얻을 수 있다. 그림 2.는 LSP 벡터를 양자화 하여 전송하는 과정을 블록 다이어그램으로 설명한다.

새로운 보코더에서는 압축률을 더욱 높이기 위해서 다음과 같은 또 다른 방법을 사용한다. AMR 7.4 kbit/s 코더에서는 고정 코드북 검색 단계에서 한 서브프레임의 여기 신호를 표현하기 위해서 서브 프레임마다 4개의 펄스를 이용하여 고정 코드북을 생성한다. 그러나 제안된 모델에서는 서브 프레임마다 오직 2개 의 펄스만을 사용함으로써 압축률을 더 높인다. 하지만 더 적은 펄스로 표현된 고정 코드북을 사용하면 복호화된 음성 신호의 질이 떨어지게 되는데, 이로 인해 떨어지는 음성 신호의 질은 앞에서 설명했던 예상하는 펄스를 적용해서 더 이상의 비트레이트의 증가 없이 더 많은 펄스를 사용함으로써 보상받을 수 있다. 그림 3.은 예상하는 펄스를 추가한 CELP 계열의 새로운 보코더의 블록 다이어그램이다.

VI. 실험 및 결과

실험에 사용된 입력 테이터 베이스는 한 여자의 음성 신호로 구성된 578개의 음성 파일이 사용되었고 모든 입력 음성 파일에서 39601개의 10차로 구성된 LSP벡터를 추출하였다. 추출된 39601개의 벡터에서 CNN 알고리즘을 사용해 1024개의 10차로 이루어진 LSP 코드북을 생성한다. 1024개의 코드북을 사용한 이유는 표 1에 있는 MOS(Mean Opinion Score)로 음질의 저하를 최소한으로 줄이기 위하여 이었다.

AMR 7.4 kbit/s모드에서는 LSP 양자화를 위해 26bit를 사용하는 반면 제안된 보코더 모델에서는 암호화 하는 부분에서 1024개의 코드북을 사용함으로써 10bit만으로 LSP 벡터를 표현 할 수 있다. 그리고 제안된 모델에서는 고정 코드북을 표현하기 위해서 2개의 펄스만을 사용기 때문에 기존 AMR 7.4kbit/s 모드 보다 서브프레임 당 6bit씩 덜 사용하며, 이는 한 프레임에서 새로운 보코더가 24bit를 더 압축할 수 있음을 의미한다. 그러므로 AMR 7.4Kbit/s코더는 한 프레임을 전송하기 위해 148bit가 필요하지만 두 가지 방법으로 접근한 새로운 보코더는 108bit만을 사용함으로써 27% 더 압축률을 높일 수 있다. 표 2

표 1. 코드북 사이즈에 다른 합성된 신호의 MOS 테스트 결과

코드북 크기	MOS 테스트 점수
128	3.45
256	3.52
512	3.62
1024	3.69

표 2. AMR 7.4 kbit/s 모드와 새로운 보코더의 한 프레임 당 비트 할당 비교

Mode	Parameter	1st sub-frame	2nd sub-frame	3rd sub-frame	4th sub-frame	total per frame
AMR 7.40 kbit/s	LSP set					26
	Pitch delay	8	5	8	5	26
	fixed code	17	17	17	17	68
	Gains	7	7	7	7	28
	Total					148
새로운 보코더	LSP set					7
	Pitch delay	8	5	8	5	26
	fixed code	11	11	11	11	44
	Gains	7	7	7	7	28
	Total					108

표 3. 여러 코더의 MOS 테스트 결과

	MOS 테스트 점수
원음	5.0
AMR 7.4 kbit/s	3.83
AMR 4.75 kbit/s	3.50
제안된 코더	3.79

는 두 코더의 비트 할당량을 나타낸 것이다.

복호화 된 음성 신호의 질을 측정하기 위해서 10명의 청자의 MOS테스트를 수행 하였다. 입력 데이터 베이스에서 한 명의 여성의 화자에 의해 얻어진 20개의 음성 파일을 랜덤으로 추출해 테스트는 이루어 졌다. 표 3은 AMR 7.4 kbit/s모드와 4.75kbit/s 모드, 그리고 새로운 코더의 MOS 테스트 결과를 비교하여 보이고 있다.

참 고 문 헌

[1] M. Decina and G. Modena, "CCITT standards on digital speech processing," IEEE Journal on Selected Areas in Communication, 6, pp.227-234, 1988.

[2] ISO/IEC 14496-3, information technology - very low bit rate audio-visual coding, part 3: Audio, Subpart 1-3, 1998.

[3] ITU-T Recommendation G.723.1, Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, 1996.

[4] C. Laflamme, J.P. Adoul, H.Y. Su, and S. Morissette, "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes," Proc. IEEE ICASSP, Vol.1, pp.177-180, 1990.

[5] ITU-T Recommendation G.729. "Coding of speech at 8kbit/s using conjugate structure algebraic-code-excited linear prediction (CS-ACELP), 1996.

[6] ETSI, Digital cellular telecommunications system(phase2): Enhanced full rate(EFR) speech transcoding (GSM 06.60 version 6.0.0), ETSI EN pp.300-726, 1997.

[7] H.J. Kim, D.G. Jee, M.H. Park, B.S. Yoon, and S.I. Choi, "The real-time implementations of AMR codec for IMT-2000 system," Advanced Communication Technology, ICACT, The 7th International Conference,(1), pp.362-365, 2005.

[8] J. Srinonchat, S. Danaher, and A. Murray, "Address vector quantisation applied to speech coding." Proc. IEEE Int. Symp. on Sig. Proc. and Info. Tech., pp.745-748, 2003.

[9] M. Schroeder and B. Ata, "Code-Excited Linear Predictive (CELP): high quality speech at very bit rate." Proc. IEEE ICASSP, pp.937-940, 1985.

[10] 김경민, 윤성완, 최용수, 박영철, 최용수, 윤대희, 강태익, "이중 전송률(2.4/4.0 kbps)을 갖는 개선된 하모닉-CELP 음성부호화기," 한국통신학회 논문지, 28권 제3C호 pp.457-462, 2003.

[11] C.H. Lee, S.K. Jung, and H.G. Kang, "Applying a Speaker-Dependent SpeechCompression Technique to Concatenative TTS Synthesizers", IEEE Trans. Speech and Audio Proc.. Vol.15, pp.632-640, 2007.

[12] 안병호, 유지상, 이승훈, 김상훈, " TTS를 이용한 멀티미디어 서비스", 한국통신학회지, 제16권 제5호, pp.534-543, 1999.

[13] 3GPP TS 26.071 V 7.0.0, "Adaptive Multi-Rate speech processing functions; General description",1999.

[14] 3GPP TS 26.090 V7.0.0, "Adaptive Multi-Rate speech transcoding", 1999

[15] 3GPP TS 26.073 V7.0.0, "Adaptive Multi Rate (AMR) speech; ANSI-C code for the AMR speech codec", 1999.

[16] D.C. Park, "Centroid Neural Network for Unsupervised Competitive Learning." IEEE Trans. Neural Networks, Vol.11, pp.520-528, 2000.

[17] D.C. Park and Y.J. Woo, "Weighted centroid neural network for edge reserving image compression." IEEE Trans. Neural Networks, Vol.12, pp.1134-1146, 2001.

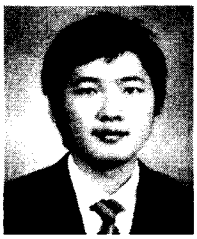
[18] 이송재, 박동철 "Bhattacharyya 커널을 적용한 Centroid Neural Network." 한국통신학회 논문지, 32권 9호, pp.861-866, 2008.

[19] Kohonen, T. "The 'neural' phonetic typewriter". IEEE Computer, 21, pp.11-22, 1988.

[20] Lei Zhang, Tian Wang and Cuperman. V. A "CELP variable rate speech codec with low average rate." IEEE International Conference on ICASSP, 2, pp.735-738, 1997.

[21] D.C. Park, O.-H. Kwon, and J. Chung, "Centroid Neural NetworkWith a Divergence Measure for GPDF Data Clustering," IEEE Trans. Neural Networks, Vol.19, No.6, pp.948-957, 2008.

민 병 제 (Byung-Jae Min) 정회원



2006년 2월 명지대학교 정보공학(공학사)
2006년~2008년 명지대학교 정보공학(공학석사)
2008년 (주)SFA Engineering
2008년~현재 LG. Philips LCD 연구원
<관심분야> 컴퓨터비전, 영상처리, 신경망, 음성처리

박 동 철 (Dong-Chul Park) 정회원

1980년 2월 서강대학교 전자공학과(공학사)
1982년 2월 한국과학기술원 전기 및 전자공학과(공학석사)
1990년 6월 Univ. of Washington (Seattle), Electrical Engineering (Ph.D.)
1990년 8월-1994년 2월 조교수, Florida Int'l Univ. Dept. of Eelct. and Comp. Eng.
1994년 3월~현재 명지대학교 정보공학과 교수
1997년~2000년 IEEE Tr. on Neural Networks, Associate Editor
1999년~현재 IEEE Senior Member
<관심분야> 신경망 알고리즘 개발, 음성인식, 멀티미디어 데이터 처리 및 분석