Neural Network을 이용한 무선 통신시스템에서의 VAD

정회원 이 호 선*, 김 수 경**, 박 승 권**

VAD By Neural Network Under Wireless Communication Systems

Hosun Lee*, Sukyung Kim**, Sung-Kwon Park** ***Regular Members***

요 약

EBF(Elliptical basis function) 신경망은 비선형 처리를 가능하게 하며, 잡음에 강하고 빠른 수렴을 하는 장점이 있다. 또한 EBF는 설계가 간단하여 실시간 음성 구간 검출기(Voice Activity Detection, VAD)에 적용하기 용이하다. 따라서 전송 효율을 높이기 위해 사용되는 음성구간 검출기를 제안함에 있어 EBF 신경망을 이용하였다. EBF의 학습 알고리즘은 K-평균 클러스터링(K-means Clustering) 알고리즘과 선형 최소 제곱 방법(Least Mean Square error, LMS)을 사용하였다. G.729 Annex B 와 RBF(Radial Basis Function) 신경망을 이용한 음성구간 검출기와 성능 비교에 있어서, G.729 Annex B 음성 검출기보다 70% 이상의 높은 성능개선을 나타냈고, RBF 신경망을 이용한 음성구간 검출기 보다 비음성 구간에서 50%정도의 높은 효율을 보였다.

Key Words : VAD, Voice Activity Detection, Neural Network, EBF

ABSTRACT

Elliptical basis function (EBF) neural network works stably under high-level background noise environment and makes the nonlinear processing possible. It can be adapted real time VAD with simple design. This paper introduces VAD implementation using EBF and the experimental results show that EBF VAD outperforms G729 Annex B and RBF neural networks. The best error rates achieved by the EBF networks were improved more than 70% in speech and 50% in silence while that achieved by G.729 Annex B and RBF networks respectively.

Ⅰ. 서 론

가변적인 음성압축 코딩에 적합한 다중접속 기술의 출현으로 셀룰러 네트워크의 차세대 다중접속 시스템상에서 대용량을 만들기 위한 가변적 음성압축 코딩은 결정적인 시스템 요소이다[1]. 음성을 전송하는 양방향 통신의 경우, 모든 순간에 데이터가 전송되는 것은 아니다. 어느 순간에는 말을 하고 또 어떤 순간에는 상대의 말을 듣기만 할 때도 있다. 통계적으로 전체 통화 구간에서 실제 음성을 전송하는 구간은 약 40%이고 나머지 60%는 침묵시간이다. 따라서 음성을 전송하지 않는 구간에서는 음성대신 다른 데이터를 전송할 수도 있다. 음성구간 검출기술은 유무선 통신에서 대역폭을 줄일 뿐 아니라, 대역을 재사용 가능하게 하고 전송 효율을 높인다. 이러한 음성 구간 검출기는 거의 모든 응용분야에서 사용되고 있으며, 음성 구간 검출기만의 국제 표준은 없으나, ITU-T의 G.729 Annex B와 G.723 Annex A 그리고 GSM의 음성 압축기술 인 RPE-LTP등에서 음성구간 검출기를 기반으로 침묵 압축기법(Silence Compression)을 표준화 하여 널리 사용되고 있다.

음성 구간 검출기에 적용할 신경망은 비선형 처리를 가능하게 해주며, 분류기(Classifier)로서의 역

* 한양사이버대학교 정보통신공학과 (r10394@ihanyang.ac.kr)
** 한양대학교 전자통신컴퓨터공학부 (sp2996@hanyang.ac.kr)
논문번호 : KICS2005-06-250, 접수일자 : 2005년 6월 20일

할을 갖는다[2]. 또한, 신경망은 설계가 간단하고 적용하기가 용이하며 잡음에 영향을 덜 받는 장점이 있다[3]. 신경회로망 중에서 RBF(Radial Basis Function) 신경망은 여러 유형의 패턴인식 문제를 다루기 위해 사용해왔다[4][5]. RBF신경망은 하나의 은닉층을 가지고 있고, 은닉층 각각의 노드(Node)는 중심점(Center)을 중심으로 방사형(Radial)으로 대칭인 응답을 출력한다. 노드 각각의 중심점은 입력 공간상에 존재하는 벡터이며 각 노드의 입력과 출력간의 함수는 비선형적이다. 최종출력은 은닉층에 있는 각 노드의 출력들에 가중치를 곱하여 더한 값이 된다.

전통적인 RBF(Radial Basis Function) 신경망 학습 방식은 입력 데이터의 특징을 이용해서 미리 은닉 뉴런의 개수와 그것의 파라미터를 고정시킨 후, 은닉 뉴런과 출력 사이의 연결 가중치는 선형 최소 제곱 방법 (Least Mean Square error, LMS)을 이용해서 추정한다. 이러한 방법을 이용하면, 처음에 문제에 적합한 은닉 뉴런의 개수를 결정하는 일이 쉽지 않을 뿐만 아니라, 새로운 데이터가 계속해서 들어오는 경우에 새로운 신경망의 크기와 파라미터들을 새로 결정해야 하기 때문에 순차적 학습 또는 온라인 학습에 사용될 수 없다는 단점이 있다. 이러한 단점을 극복하기 위해, 새로운 데이터가 입력 될 때마다 은닉 뉴런의 파라미터 및 가중치의 값이 실시간으로 학습하도록 설계하였다. 또한 RBF 신경망은 많은 수의 은닉층을 요구할 때가 많아 복잡도를 증가시키며 중복된 입력 자료에 민감하다. 반면 EBF신경망은 첫번째 은닝층에서 입력층의 차원을 축소 시킴으로써 RBF 신경망이 가지는 문제점을 보완할 수 있다. 본 논문에서는 EBF 신경망을 이용한 음성구간 검출기를 설계하였고, 은닉층은 K- 평균 클러스터링(K-means Clustering) 알고리즘을 은닉 뉴런과 출력 사이의 연결 가중치는 LMS 학습 알고리즘을 사용하여 RBF 신경망과 G.729 Annex B와의 성능을 각각 비교, 분석하였다.

Ⅱ. 음성 구간 검출기

어떤 음성 파형 중 각 구간이 음성인지 비음성인지를 판별하기 위해서는 그림 1과 같은 과정을 걸친다. 우선 음성신호를 수신하여, 수신된 신호의 분석을 통해 특정 파라미터를 추출한다. 이 파라미터들은 EBF 신경망의 입력값이 되고, EBF 신경망을 걸쳐 나온 출력 값을 근거로 음성 또는 비음성으로 판단한다.

그림 1. 음성구간 검출기 순서도

2.1 신호 분석

신경 회로망의 입력 변수는 단구간 전력, 피치의 안정성, 스펙트럼 포물선 등에서 사용된다. 이 값들은 음성/비음성 구간을 결정하는데 최적의 음성 파라미터이다[3]. 이러한 파라미터 값들은 신경망의 입력 값이 되어 학습 및 검증을 통해 스칼라의 결과값을 얻고, 이 값은 임계 값을 통해 음성/비음성 구간 여부를 결정짓는다.

첫 번째 파라미터, 단구간 평균전력(Short-time Average Power)은 자기 상관 함수의 대수 값이다. 이 값은 비음성 구간에서 작고, 음성구간에서 큰 값을 갖는다.

$$E = 10 \log[\frac{1}{N} \sum_{n=m-N+1}^{m} S_w{}^2(n)] \quad (1)$$

두 번째 파라미터는 피치의 길이차(Pitch Period Difference)이다. 현재 구간과 이전 구간 사이의 피치 길이 차이는 피치의 안정도와 관련이 있으며 안정적일 때 큰 값을 갖는다. 이 값은 G.729의 음성구간 검출기에도 사용된다.

$$P = \max \left[\frac{\log \sum_{t=0}^{N-1} \{r(t)r(t-\tau)\}}{\log \sum_{t=0}^{N-1} \{r(t)r(t)\}} \right] \quad (2)$$

이때 각 샘플당 τ 는 $20 \le \tau \le 160$ 이고, r(t) 는 시간 t 에서의 선형 예측 오차 신호이다. $t = 0$ 일 때는 분석할 구간의 시작점 이다.

마지막 파라미터는 스펙트럼 포물선의 특징과 관련이 있는 영차 ML 파라미터(Zero-order Most Likely Parameter) 이다.

$$F = \log \sum_{i=0}^{p} \alpha_i^2 \quad (3)$$

이때 α_i 는 선형 예측 계수, p 는 선형 예측 분석의 차수이다. 이 값은 음성 스펙트럼 포물선과 평면(flat) 스펙트럼과의 거리와 관련이 있다.

2.2 신경망

신경망(Neural Network)은 인간 두뇌의 신경세포를 모방한 개념으로 마디(Node)와 고리(Link)로 구성된 망구조를 모형화하고, 과거에 수집된 데이터로부터 반복적인 학습과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법이다. 신경망의 가치는 불완전하고 잡음이 많은 입력의 해석뿐만 아니라 패턴인식(Pattern recognition), 학습, 분류, 일반화 등을 위한 활용성에 있다. 신경망은 전문가 시스템의 논리적이고 분석적인 기법을 활용해서도 시뮬레이션 하기 어려운 인간의 문제해결의 지원할 수 있다. 신경망의 여러 알고리즘 중 EBF 네트워크를 음성 검출에 이용하고자 한다. 대표적인 신경망 중 하나인 RBF는 계산과정이 간단하고 시간이 작게 걸린다는 장점이 있으나, 수렴성에 있어 국부최소값(Local minimum)에 오류를 범할 가능성이 많다. 특히 중심 값을 잘못 잡거나 은닉마디의 수가 충분하지 않은 경우에 이러한 현상이 발생한다. 따라서, RBF 신경망은, 많은 수의 은닉마디를 요구할 때가 많으며, 필요한 은닉마디의 수는 변수가 많아질수록 급속하게 복잡도가 증가할 뿐 아니라, 불필요하거나 중복된 입력자료에 대하여 민감하다. 반면 EBF 신경망을 이용하면 첫 번째 은닉 층에서 입력층의 차원을 축소시킴으로써 RBF 신경망이 가지는 문제점을 보완할 수 있다.

EBF 신경망은 RBF신경망의 확장형 모델로서 고려될 수 있다[6][7]. EBF 신경망에서 N 개의 입력과 M 개의 중심 함수를 갖을 때 k 번째 출력 값은 다음과 같다.

$$y_k(\vec{x}_p) = w_{k0} + \sum_{j=1}^{M} w_{kj}\phi_j(\vec{x}_p) \quad (4)$$

$p = 1, \Lambda\ , N$, $k = 1, \Lambda\ , K$ and $j = 1, \Lambda\ , M$

활성함수 $\phi_j(\cdot)$ 는 j 번째 베이시스함수로, 입력 벡터와 중심점 사이의 거리를 계산한다.

$$\phi_j(\vec{x}_p) = \exp\left\{-\frac{1}{2\gamma_j}(\vec{x}_p - \vec{\mu}_j)^T \sum_j^{-1}(\vec{x}_p - \vec{\mu}_j)\right\} \quad (5)$$

식 (4)와 식 (5)에서 $\vec{x}_p$ 는 p 번째 입력 벡터이고, μ_j 는 평균 벡터, $\sum_j$ 는 j 번째 베이시스 함수의 공분산(covariance) 행렬이다. ω_{k0} 는 바이어스값, ω_{kj} 는 각 j 번째 베이시스함수의 가중치 값이다. γ_j 는 j 번째 베이시스함수의 값을 조절하는 파라미터로

$$r_j = \frac{3}{5}\sum_{k=1}^{5}\left\|\vec{\mu}_k - \vec{\mu}_j\right\| \quad (6)$$

으로 나타내고, $\vec{\mu}_k$는 유클리드 값에 따라 $\vec{\mu}_j$ 의 k 번째 인접 값에 해당한다. 이것은 또 다른 신경망 RBF에서 함수의 폭 결정에 사용되는 K -nearest neighbor 알고리즘과 유사하며 실험을 통해 5개의 이웃 함수의 중심점을 사용하고, 평균 거리에 3을 곱하는 것이 합리적인 결과임이 나타났다[8]. EBF 신경망 설계에 있어 주요 사항은 각 은닉마디 함수의 중심점과 출력 값에 곱해지는 가중치 값을 결정하는 것이다. 이 값들은 고정된 것이 아니라, 각 입력 시에 따라 실시간으로 학습하여 최적 값을 나타내도록 했다. 학습에 사용되는 알고리즘으로 LMS (Least Mean Square)와 K -평균 클러스터링 알고리즘을 사용하였다. 신경망의 입력 값들은 벡터 공간상에서 특정지역에 무리 지어 분포하므로 베이시스 함수의 중심 값은 이들 무리의 중심으로 결정하여 입력 값을 분류한다. 우선 각 함수의 중심점 계산으로는 K -평균 클러스터링(K-means Clustering) 알고리즘과 샘플 공분산 (Sample Covariance)값을 사용했다. 평균 벡터를 결정하기 위해 입력 값의 집합 $\chi^{(k)}$ 를 j 개의 은닉마디 함수의 개수에 따라 $J^{(k)}$ 의 집합으로 나눈다. 그리고 나서 각 집합에서의 평균 값을 함수의 중심인 $\vec{\mu}_j$ 값으로 결정한다.

$$\vec{\mu}_j \approx \hat{\vec{\mu}}_j = \frac{1}{N}\sum_{\vec{x}\in\chi_j}\vec{x} \quad (7)$$

이때 $\vec{x} \in \chi_j$ 이고, $\forall j \neq k, N_j$ 에서 $\left\|\vec{x} - \vec{\mu}_j\right\| < \left\|\vec{x} - \vec{\mu}_k\right\|$ 이면, N_j 클러스터 χ_j 의 샘플이다. 공분산 행렬 값은 대략 샘플 공분산 값으로 결정된다.

$$\Sigma_j \approx \hat{\Sigma}_j = \frac{1}{N_j}\sum_{\vec{x}\in\chi_j}(\vec{x} - \vec{\mu}_j)(\vec{x} - \vec{\mu}_j)^T \quad (8)$$

중심 값과 공분산 값은 새로운 입력 값 $\vec{x}_p$ 에 따라 실시간으로 갱신된다. 가중치 값은 기존의 LMS방

식을 사용한다. j 번째 원하는 출력에서 EBF신경망의 j 번째 출력 값을 뺀 오차의 식을 식(9)에 정의하면 가중치 값은 식 (10)으로 학습된다.

$$e_j(t) = d_j(t) - \hat{\Phi}^T(t) w_j(t-1) \tag{9}$$

이때 $1 \le j \le M$, $\hat{\Phi}^T(t) = [\phi_1(t) \Lambda\ \phi_k(t)]^T$ 이다.

$$w_j(t) = [w_{1j}(t) \Lambda\ w_{kj}(t)] \tag{10}$$

$$w_j(t) = w_j(t-1) + \Delta \hat{\Phi}^T(t) e_j(t) \tag{11}$$

여기서 사용된 LMS 알고리즘과 K-평균 클러스터링 알고리즘은 선형 학습 알고리즘으로 빠른 수렴속도를 갖는다. 빠른 수렴속도는 고속 통신과 같은 실시간 응용분야에서 활용하기 적합하다.

2.3 음성/비음성 결정

위의 EBF 신경망에 출력 값으로 음성/비음성을 판단하는 것 보다 신경망의 출력 값을 일반화 하면 최적의 임계 값을 찾기가 쉽고, 음성/비음성의 판단도 용이해진다. 일반화 (normalized)함수로는 다음의 함수를 사용하여 0부터 1까지의 값으로 일반화 시켰다.

$$OUT = \frac{1}{1 + e^{-IN}} \tag{12}$$

Ⅲ. 실험 및 성능분석

3.1 실험 환경

실험은 다음과 같은 과정을 걸쳐 이루어 졌다. 우선, 음성 데이터 신호를 분석하고, 분석된 파라미터 값을 EBF 신경회로망의 입력값으로 사용하였다. EBF 신경회로망을 걸쳐 나온 출력값을 일정 임계값으로 구분하여 음성/비음성 여부를 판단하였다.

실험에 사용된 음성신호는 총 길이 20,000구간의 8kHz 표본화, 16bit 선형 양자화한 wave 형식의 파일이다. 신호의 크기는 -24dBoV에서 -40 dBoV 사이이며, -63, -58 53 -48 dBoV 가우시안 노이즈를 삽입하여 실제 생활과 비슷한 환경으로 설정하였다. 음성 신호 분석은 블록 크기 10, 20ms 크기의 해밍 창함수를 사용하였고, 선형예측에는 10개의 과거 샘플수를 사용했다. 신경망의 입력 층은 신호 분석을 통해 얻어진 3개의 파라미터(단구간 평균전력, 영차 ML 파라미터, 피치길이 차)를 사용했고, 은닉층 마디의 개수는 10개로 설정하였다. 은닉층에서 각 마디의 중심값 결정을 위해 K-means clustering과 Sample covariance 알고리즘을 사용하였고, 이는 새로운 음성 샘플이 수신될 때마다 실시간으로 학습한다. 은닉층의 출력은 가중치를 곱하여 한 개의 출력층으로 전달되는데, 이때 가중치는 신호처리에서 널리 사용되는 LMS 알고리즘을 학습알고리즘으로 사용했다.

EBF출력 값에 따라 음성/비음성을 구분하기 위해 임계 값을 설정해야 한다. 따라서 임계 값에 변화에 따른 오차율을 알아보았다.

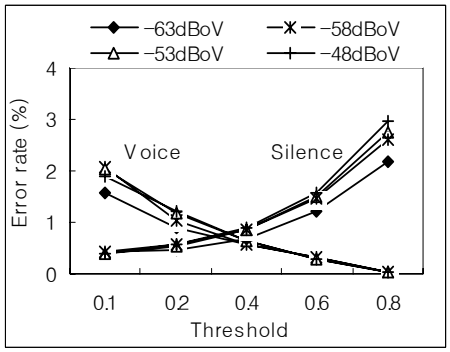

그림 2. 임계 값에 따른 오차율

그림 2는 임계값에 따라 음성구간과 비음성 구간에서 오차율을 나타낸 것이다. 실험을 통해 임계 값이 커질수록 음성구간에서는 오차율이 낮아지고 비음성구간에서는 오차율이 커지는 것을 보였다. 음성/비음성 구간 모두에서 낮은 오차율을 얻기 위한 적절한 임계 값은 0.2~0.4으로, 실험에서는 0.3으로 임계값을 설정하였다.

3.2 타 음성구간 검출기와의 성능 비교 분석

EBF 신경망을 이용한 음성구간 검출기는 ITU-T G.729 Annex B의 음성구간 검출기, 널리 사용되고 있는 신경망인 RBF 각각과의 성능을 비교하였다. 통계상으로 알려진 바와 같이 음성구간이 전체 통신 부분의 약 40%를 차지하고 있고, 나머지 60%가 비음성 구간이다. 따라서, 그림 3에서는 비음성 구간에서, 그림 4에서는 음성구간에서 각 알고리즘간의 성능 비교를 나타내었다. EBF 신경망을 사용하는 VAD는 어떤 잡음환경[9]에서도 G.729 Annex B 보다 나은 성능을 보였다.

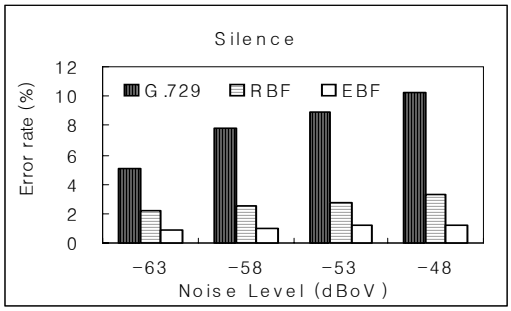

그림 3. 비음성 구간에서 EBF를 사용한 VAD와 RBF, G.729와의 성능 비교

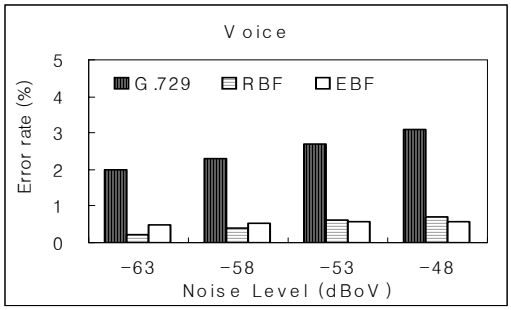

그림 4. 음성 구간에서 EBF를 사용한 VAD와 RBF, G.729와의 성능 비교

그림 3, 4에서 나타난 바와 같이, EBF 신경망을 사용한 검출기의 성능이 G.729 Annex B나 RBF보다 우수함을 볼 수 있다. 특히 EBF 음성검출기는 G.729 음성 검출기에 비해 음성/비음성구간 모두에서 70%이상의 높은 성능향상을 보였다. RBF 음성검출기와의 비교를 보면 음성구간에서의 잡음의 세기에 따라 약간의 차이를 보이지만 전반적으로 오차율은 거의 비슷하게 나타났고, 비음성 구간에서 50% 정도의 성능 향상을 보였다.

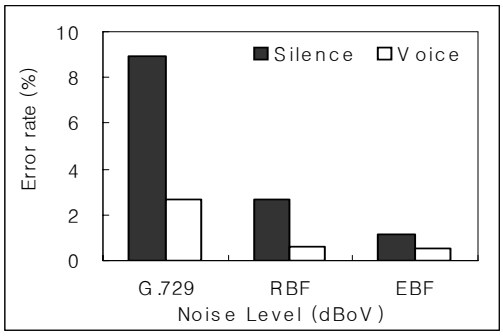

그림 5. 각 음성 검출기의 음성/비음성 구간의 오차율

또한, G.729나 RBF의 음성검출기에 비해 EBF 음성검출기는 잡음에 영향을 거의 받지 않는 것으로 분석 되었다. 즉, 잡음의 세기에 따라 1.5배~3배 정도의 오차율 차이를 보이는 다른 검출기와 달리, 잡음의 세기에 상관없이 거의 일정하게 낮은 오차율을 갖는다. 또한 그림 5에서 보이는 바와 같이, G.729와 RBF의 음성검출기는 음성구간에 비해 비음성 구간의 오차율이 현저하게 높은 반면, EBF 음성검출기는 음성구간 뿐 아니라 비음성 구간에서도 오차율이 작게 나타난다.

Ⅳ. 결 론

본 논문에서는 EBF 신경망을 이용한 음성 구간 검출기를 설계하여 그 성능을 분석해 보았다. EBF 신경망은 잡음에 강하고 구조의 단순함과 빠른 학습속도를 지녀 실시간 통신에 적용될 수 있는 강점을 지니고 있다. 또한 입력이 들어올 때마다 중심점과 가중치를 실시간으로 갱신하는 학습을 통하여, 입력에 따라 적응적으로 변할 수 있게 하였다. 실험을 통해 음성/비음성 구간의 적절한 임계값을 설정하였고, G.729 Annex B와 RBF를 이용한 음성구간 검출기와의 성능비교 및 분석을 하였다. 실험 결과 EBF 음성검출기는 잡음의 영향을 거의 받지 않으며, G.729나 RBF의 음성검출기가 음성구간에 비해 비음성구간에서는 높은 오차율을 보이는 반면, EBF의 음성검출기는 음성/비음성 구간 모두에서 낮은 오차율을 보였다. 이러한 높은 효율과 신경망이 가지는 빠른 수렴 및 학습 속도를 바탕으로, VAD는 무선통신에서뿐 아니라 VoIP(Voice over Internet Protocol)와 같은 유선통신 분야에도 적용될 수 있다.

참 고 문 헌

[1] Gersho A. and Paksoy E. "An Overview of Variable Rate Speech Coding for Cellular Networks", *IEEE Conf. Selected on Topics Wireless Commun*, Vancouver, pp.172-175. 1992,

[2] Jacek M. Zurada, "Introduction to Artificial Neural Systems", *West Publishing Company* , 1992

[3] Ikedo, J. "Voice Activity Detection Using Neural Network", *IEICE Trans. Commun.*, Vol. E81-B, No. 12, pp.2209-2513, 1998

[4] S.Renals, “Radial basis function for speech pattern classification,” *Electron. Lett.*, vol. 25, no.7, pp.437-439, 1989.

[5] Y.Lee, “Handwritten digit recognition using K-nearest-neighbor, radial basis function, and back propagation networks”, *Neural computing*, vol. 3, no.3, pp. 440-449, 1991

[6] J. Moody and C.J. Darken, “Fast learning in networks of locally tuned processing units,” *Neural Comput.*, vol. 1, pp.281-194,1989.

[7] D.S.broomhead and D. Lowe, “Multivariable function interpolation and adaptive networks,” *Complex Syst.*, vol.2, pp.321-355, 1988.

[8] Man-Wai Mak, “Estimation of Elliptical Basis Function Parameters by th EM Algorithm with Application to Speaker Verification”, *IEEE Trans. Neural network.*, Vol.11, pp.961-969, 2000.

[9] A. Benyassine, E. Shlomot, and H-Y. su, “ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application”, *IEEE Commu. Mag.*, vol.35, no.9, pp.64-73, 1997.

이 호 선 (Hosun Lee) 정회원

1983년 2월 한양대학교 전자통신공학과 공학사

1988년 8월 미국 University of Missouri 공학석사

2000년 3월~현재 한양대학교 전자통신전파공학과 박사과정

1988년 10월~1991년 2월 삼성 SDS 근무

1991년 2월~2004년 12월 모토로라 코리아 근무

2005년 3월~현재 경문대학, 한양대학교, 한양사이버대학교 강사

<관심분야> 디지털 신호처리, 이동통신공학, 반도체 공학

김 수 경 (Sukyung Kim) 정회원

2002년 2월 명지대학교 전자정보통신공학과 공학사

2002년 3월 한양대학교 전자통신컴퓨터공학과 석사과정

<관심분야> 디지털 신호처리, 디지털 CATV Systems, 이동통신공학

박 승 권 (Sung-Kwon Park) 정회원

1982년 2월 한양대학교 전자통신공학과 공학사

1983년 8월 Stevens Institute of Technology, 전자공학과 공학석사

1987년 12월 Rensselaer Polytechnic Institute, 전자공학과 공학박사

1984년 1월~1987년 8월 Rensselaer Polytechnic Institute, Electrical, Computer and Systems Engineering Dept., Research Assistant

1987년 9월~1992년 8월 Tennessee Technological University, Electrical Engineering Dept., 조교수

1992년 9월~1993년 1월 Tennessee Technological University, Electrical Engineering Dept., 부교수

1993년 3월~현재 한양대학교 공과대학 전자전기컴퓨터공학부, 교수

<관심분야> 지능형 데이터 방송, CATV Multimedia Systems, Digital Signal Processing