《主 題》

신경망을 이용한 음성인식 시스템

석 용 호* · 김 기 철* · 한 일 송** · 이 황 수*
(*한국과학기술원 정보및 통신공학과, **한국통신 연구개발원)

■ 차 례 ■

요 약

본 글에서는 음성인식에 적용된 신경망 구조를 알아 본다. 또한 신경망 VLSI와 국내에서 개발된 신경망 VLSI인 URAN에 대해서 살펴보고 URAN을 이용한 음성인식 시스템의 설계에 관해 기술한다. 시뮬레이션을 통해 낮은 정밀도의 입출력 및 연결강도, 선형 출력함수를 가지는 뉴런을 사용하는 신경망 음성인식 시스템의 성능을 분석하고 잡음 환경에서 낮은 정밀도를 사용한 신경망의 성능저하 정도를 검토한다.

Ⅰ. 서 론

음성인식은 신호처리, 음향학, 음운학, 언어학 등을 포괄하는 종합학문이며 우리생활에 커다란 편의를 줄 수 있는 지식 집약적 기술이다. 그러나 현재까지의 음성인식 시스템은 실용적인 관점에서 아직 만족할 만한 수준의 성능을 보이지 못하고 있는 실정이다. 따라서 이러한 약점을 보완하기 위한 연구가 다각적으로 진행되고 있다. 음성인식 방식에는 크게 음성특징 패턴의 시간축에서의 비선형적 정합을 이용하는 패턴 정합 방식, 음성의 통계적인 성질을 조사, 인식하는 통계학적 방식, 음성자체의 여러 성질에 대한 지식을 정리, 종합하여 대상 음성을 인식하는 지식 기반 방식과 인간의 두뇌에서의 신경세포의 동작을 흉내내는 신경망을 이용하는 방식 등이 있다. 본 글에서는 그중 인간의 두뇌를 흉내낸 신경망을 이용한 방식에 대하서 알아본다.

신경망을 이용한 음성인식 방식의 장점은 입력을 이용해서 스스로 학습할 수 있으며 음성 신호에 내재된 특징을 자연스럽게 추출할 수 있다는 점이다. 더우기 목표 음성과 다른 음성의 차이점을 스스로 학습하게 되므로 자연스럽게 변별력을 가지게 된다. 또한 몇 개의 제한점을 동시에 최대한 만족시키도록 학습될 수 있으며 입력, 출력에 대한 통계적 가정을 하지 않아도 되므로 잘못된 가정에 의한 오차를 방지할 수 있다. 가장 중요한 장점은 간단한 단위 구조의 반복으로 전체 시스템을 구현할 수 있으며, 따라서 내부 결손과 잡음에 강인하다는 점이다[1][2][3].

음성인식에 이용되는 신경망은 크게 정적 신경망(Temporally Static Neural Network)과 동적 신경망(Temporally Dynamic Neural Network)으로 나뉘게 된다. 또한 신경망의 특성상 음소, 단어 레벨의 경우에는 통계적 방식 또는 시간 정합방식 등과 대등하거나 더 좋은 성능을 보이지만, 전체 시스템 구현상의 문제

점을 감안해서 신경망과 기존의 음성인식 알고리즘을 결합하려는 접근 방식이 있다.

신경망의 대규모 병렬성을 효율적으로 적용하기 위해 여러가지 하드웨어 구현 방법이 연구되고 있으며, 아날로그 방식, 디지탈 방식, 아날로그/디지탈 혼합 방식 등으로 신경망 VLSI 구현이 이루어지고 있다. 신경망 VLSI의 경우 대규모 용량의 연결고리를 구현하기 위해, 낮은 정밀도를 사용하는 경우가 많다. 역전파 신경망의 연결고리 강도의 경우 일반적으로 학습시엔 16 bit, 학습이 이루어진 뒤 테스트한 시엔 8 bit 정도의 정밀도가 필요하다고 한다[4]. 본논문에서는 범용의 아날로그/디지탈 혼합방식 신경망칩을 이용한 실시간 음성인식 시스템을 실제하고, 이의 구현시 고려되어야 할 낮은 정밀도 계산을 통해 역전파 신경망의 성능을 검토한다.

제 II 절에서는 일반적인 음성인식 시스템과 음성인식에 이용되는 여러가지 방식의 신경망에 대해서 서술하였다. 제 III 절에서는 신경망의 VLSI 구현에 대해서 알아 보며 아날로그/디지탈 혼합 방식의 신경망 VLSI중 Universally Reconstructable Artificial Neural Network(URAN)에 자세히 설명하였다. 제 IV 절에서는 URAN 및 이를 이용한 음성인식 시스템의 실제에 대해서 기술한다. 제 V 절에서는 낮은 정밀도 계산에 의한 시뮬레이션을 통해 잡음이 존재하는 환경하에서의 신경망의 숫자음 인식 성능을 실험하고 그 결과를 분석하였다. 제 VI 절에서는 결론으로 신경망을 이용한 음성인식 시스템의 성능향상에 필요한 연구 방향을 고찰하였다.

II. 음성인식을 위한 신경망

음성인식의 과정은 먼저 음성신호를 전기적 신호로 바꾸는 것에서부터 시작된다. 인간의 가청 주파수는 20 Hz-20 kHz 이지만 음성신호의 정보는 거의 1 kHz내에 집중되어 있으므로 보통 8 kHz 16 kHz정도로 샘플링하며 이때 8 bit에서 16 bit가량의 정밀도가 요구된다. 음성신호의 에너지와 영교차율 등을 이용해서 각 단어의 끝점을 검출하게 되는데 이때 무발성 구간과 배경잡음의 구분이 커다란 난점으로 작용한다. 그후 특징추출 과정이 이어진다. 음성신호의 특징추출과정은 음성인식 시스템의 성능과 직결되며, 다른 단어간의 차이를 잘 표현하는 한편, 같은 낱어에 대해서는 화자의 상태에 따른 변화에 둔감한 특징을 추출해야 고성능의 음성인식 시스템 구현이 가능하게 된다. 보통 선형 예측(Linear Predictive Coding, LPC) 계수, cepstral 계수, mel cepstral 계수, filter bank output 등을 이용한다. 이렇게 구한 특징 패턴을 학습과정중 생성된 음소/낱어의 특징패턴과 비교하여 인식낱어를 결정한다. 비교 방식은 시간 왜곡 정합 방식, 통계적 방식, 지식 기반 방식, 신경망을 이용한 방법 등이 있다. 후 각종 언어처리를 통해서 인식문장을 결정하게 된다.

신경망은 우선 인간의 신경세포를 간단하게 모델링한 것으로 각 모델을 인공신경망 위하는 시스템을 구현하게 된다. 신경망의 장점은 대규모의 병렬처리 구조에 있다. 또한 자기 학습기능이 있으며 주어진 데이타의 숨은 정보를 추출해 낼 수 있다. 음성인식에 이용되는 신경망은 입력 음성신호 또는 특징벡터의 시간축상의 변이(time shift)에 대한 고려 방식에 따라 정적 신경망(neural network)과 동적 신경망으로 나뉘게 된다. 또한 신경망과 기존의 음성인식 알고리즘을 결합하려는 접근 방식이 있으며, 연속음성을 인식하기 위해 음성인식 알고리즘의 시간 정렬(time alignment) 과정을 신경망으로 구현하려는 구조도 연구되고 있다.

2.1 정적 신경망을 이용한 음성인식

정적 신경망을 이용한 음성인식 시스템은 입력 음성을 정적인 패턴의 연속으로 보고 다층 퍼셉트론(Multi Layer Perceptron, MLP), Self organizing feature maps, Learning Vector Quantization(LVQ) 등을 이용해서 분류하는 구조이다. 이 방법은 음성신호 또는 특징벡터의 작은 시간 변이에도 영향을 받기 때문에 미리 각 음소/낱어간의 정밀한 정렬을 해줘야 한다. 시간변이의 영향을 받지 않도록 입력층의 노드갯수를 충분히 크게 하여, 시간 정렬과정을 공간적으로 펼친 것처럼 처리하거나 입력의 시간적인 context를 함께 처리함으로써 좀더 향상된 성능을 얻을 수 있으나[5] 출력에 대한 후처리과정이 없다면 그 적용범위가 단어의 인식에 제한될 수 밖에 없다. 정적 신경망의 구조는 비교적 간단하며 학습알고리즘도 안정되어 있다. 보통 낱독시스템으로 사용되기도 하지만 전체 시스템의 일부로서 이용되는 경우도 있다. 이럴경우 정적 신경망은 주로 버전형 분류기로 이용된다.

Huang과 Lippman은 3층의 역전파 신경망을 이용한 모음 분류기를 구성하였다. 입력은 2개의 포만트 주파수 F1와 F2이며 출력은 해당 모음이다. 학습은 오차 역전파를 이용해서 수행되었으며 모음 구분에 만

촉할 만한 결과를 보였다고 한다. 또한 Lippman과 Gold는 고립단어 인식에 3층의 MLP를 이용하였다[6]. 12 kHz로 샘플링된 음성에 대해 15차의 mel-scale cepstra를 10ms의 간격으로 구해서 입력으로 사용하였으며 인식률은 92.3%에 달했다. Elman과 Zipser는 실제 음성을 입력으로 해서 /b, d, g/의 음소를 분류하였다. 학습은 역시 오차 역전파 방식으로 이루어졌다. Peeling과 Moore역시 3층의 MLP를 이용해서 고립 숫자음 인식을 행하였다. 50개의 은닉 노드를 사용하였고 60쌍의 16차 세수 cepstra를 입력으로 이용하였다. 이러한 조건하에서 화자 종속인 경우에는 99.7%의 인식률을 보였다. 또한 radial basis function(RBF)를 이용한 분류방법은 MLP를 이용할때보다 좀 더 좋은 성능을 보였다. 이때 radial basis function은 보통 multi variable Gaussian function을 이용한다[7].

Kohonen은 Self-organization feature map[8]을 이용해서 neural phonetic typewriter를 구현하였으며 이를 supervised training방식으로 변형한 LVQ를 고안하였다[9]. 신경망을 이용해서 입력 특징 Vector를 동적으로 양자화하는 알고리즘이다. LVQ의 장점은 보통의 Vector 양자화(Vector Quantization, VQ) 알고리즘과는 달리 supervised training 방식이며 또한 기존의 정적 VQ에 비해서 좀더 좋은 성능을 보인다는 점이다. 따라서 기존의 정적 VQ를 대신하게 되면 전체 시스템의 성능이 향상된다.

2.2 동적 신경망을 이용한 음성인식

동적 신경망을 이용한 음성인식 시스템은 입력 음성 특징을 시간에 따라 변화하는 동적인 패턴으로 보며 또한 입력 음성 특징의 시간 지연을 고려해서 처리하는 인식 시스템을 말한다. 따라서 음성신호 또는 특징벡터의 시간축상의 변이에 영향을 적게 받지만, 신경망 내부에서 처리될 수 있는 동적인 패턴의 단위에 대한 segmentation 과정이나 안정된 학습을 위한 학습 알고리즘의 변형 등이 고려되어야 한다. 동적 신경망에는 크게 시간지연 신경망(Time Delay Neural Network, TDNN)과 재귀 신경망(Recurrent Neural Network)으로 나뉜다.

TDNN은 Waibel에 의해 발표되었으며 각 층의 출력의 시간지연된 값들이 다음층의 입력으로 들어가게 된다[10]. 시간 불변성은 중간 노드 출력의 시간지연값들과 그에 해당되는 가중치에 의해서 이루어지게 된다. 그 결과 시간축의 위치에 대한 정보는 제거되며 시간 변이에 상관없이 입력의 특징만을 구할 수 있게 된다. 따라서 인식 음소/단어의 정밀한 시간 정렬이 불필요하게 된다. TDNN의 또다른 장점중의 하나는 TDNN의 은닉층의 출력값을 조사함에 따라 각 입력 음성의 유성음/무성음과 같은 특징이 은닉층에 나타나게 된다는 것이다. 학습은 변형된 오차 역전파 방식에 의해서 학습되어 진다. P. Haffner[11]와 H. Hild and A. Waibel[12]은 Multi-State Time Delay Neural Network(MS-TDNN)을 제안하였는데, MS-TDNN은 TDNN의 시간 불변성과 Dynamic Time Wraping (DTW)의 비선형 시간축 왜곡 정합을 결합한 모델이며 단어 인식에 높은 성능을 보인다. MS-TDNN은 먼저 TDNN에 음성특징을 입력시킨 뒤 TDNN의 출력이 음소가 되도록 학습시킨다. 그후 출력 인식 음소에 DTW를 적용해서 최종적으로 단어를 인식한다. 이러한 MS-TDNN은 연속음성 인식에도 시도되고 있다.

재귀 신경망은 각 노드의 출력이 다시 바로 전층이나 그 이전층의 입력으로 사용되는 신경망 구조를 말하며 Watrous에 의해서 시간 변이 불변 음소인식에 재귀 신경망이 처음으로 이용되었다[13]. 재귀 신경망은 작은 인식단위나 제한된 인식구문에 대해서는 비교적 좋은 성능을 보이지만 학습과 분석, 설계가 아주 어려우며 학습 시간 역시 아주 길게 요구된다. 이러한 학습의 어려움 때문에 많이 이용되지 않았지만 변형된 오차 역전파 학습방식이 개발됨으로 인해서 연구가 다시 활발해 지고 있다.

예측 신경망(Predictive Neural Net, PNN)은 신경망을 비선형 예측기로 학습시킨 후 예측 오차를 분석해서 음성을 인식하는 구조이다[14]. 각 단어마다 신경망을 이용한 예측기가 할당된다. 그후 예측된 프레임값과 실제의 프레임값과의 거리를 예측 오차 또는 왜곡으로 간주할 수 있다. 따라서 인식 단어는 입력음성 신호와의 누적 예측 오차가 가장 작은 예측 모델의 단어로 결정된다. 또한 시간에 따른 최적의 예측 모델을 결정하기 위해서 PNN의 출력과 실제 음성 특징과의 차이를 cost로 하는 동적 시간 왜곡 정합 기법이 적용되기도 하며, 이를 이용한 연속음성 인식이 연구되고 있다.

은닉 제어 신경망(Hidden Control Neural Network)에서는 PNN과는 달리 오직 하나의 예측기만 존재하고 각 단어 모델마다 제어 입력이 별도로 존재하게 된다. 따라서 입력 음성에 대해서 모든 단어 모델의 제어 입력을 차례로 예측기에 입력한 뒤 예측 오차가 가장 작은 제어 입력에 해당되는 단어 모델을 DTW 방식에 의해서 구한다[5].

2.3 신경망과 기존의 음성인식 알고리즘의 결합

신경망으로 구성된 음성인식 시스템의 약점은 계층적 구성이 곤란하고 음성의 시간적 흐름 특성을 잘 모델링 할 수 없다는 점이다. 따라서 계층적 구성이 가능하고 시간적 특성을 잘 모델링해 주는 기존의 음성인식 알고리즘과의 결합에 대한 연구가 수행되어 있다. 그중에서도 주로 신경망과 DTW 또는 Hidden Markov Model(HMM)의 결합에 관한 많은 연구가 수행되어 좋은 결과를 보여주었다. 이때 신경망은 보통 기존의 왜곡 척도나 vector 양자화기를 대신한 비선형 분류기 또는 예측기로 이용되며 DTW, Viterbi alignment, 또는 HMM을 사용하여 음소/낱말에 대한 순서, 지속기간등의 시간적인 제약을 반영시키는 방식으로 처리된다.

신경망과 HMM의 결합은 여러가지 방법이 있다. 이론상으로 평균 자승 오차 억전파로 학습된 MLP의 출력값은 해당 Class에 대한 확률의 추정값으로 볼 수 있다[16][17]. 따라서 학습된 MLP의 출력값을 HMM의 관측 확률값으로 이용할 수 있다. 또한 HMM 구조에서는 각 입력에 Label을 부여함으로써 MLP의 학습을 가능하게 한다. 즉 MLP와 HMM의 융합적인 반복 최적화 학습이 이루어진다.

재귀 신경망과 HMM과의 결합도 연구되고 있으며 다음은 재귀 신경망과 HMM과의 차이점 및 상호 보완 방법을 설명한 것이다[18].

- 신경망은 변별력을 가지는 결정을 제공한다. 즉, 학습과정에서 목표 Class와 출력에 대한 거리는 최소화 시키는 동시에 다른 Class와 출력에 대한 거리는 최대화 시키게 된다. 보통의 HMM은 이러한 변별력이 약하다.
- 재귀 신경망은 단기간 context 정보를 내부적으로 가지게 된다. 따라서 출력은 context 독립 음소로 나타내어 질 수 있다. HMM은 이러한 내부 정보를 가지지 않으므로 triphone과 같은 context 정보를 별도로 고려해야 한다.
- 신경망은 내부적으로 화자 변이에 대한 적응성을 가진다. 화자의 차이에 대한 정보는 많은 학습에 의해 자동적으로 신경망에 내재된다.
- 신경망은 보통 gradient descent 방법에 의해서 학습된다. 따라서 HMM의 Baum-Welch 파라메터 재추정 학습방법에 비해 학습이 느리다.

Viterbi Net은 시간 정렬과 순차 제어의 문제를 신경망으로 해결하는 구조이다[6]. 간단한 단위 구조의 적절한 배치와 연결에 의해서 Viterbi alignment 과정을 자동적으로 수행하게 된다. 이 방법에 의한 인식률은 기존의 HMM에 의한 인식률과 거의 비슷한 성능을 보인다. Viterbi Net의 동작이 기존 HMM의 Viterbi alignment와 실제로는 동일하기 때문이다.

DTW 또는 HMM의 결과를 다시 신경망으로 처리하는 구조는 DTW 또는 HMM의 매개한 출력값을 신경망에 의해서 재분류하거나 분별 정보 등을 신경망으로 구현해서 인식률을 향상시키는 구조 등이 있다[19].

III. 신경망의 VLSI 구현

3.1 신경망 VLSI와 하드웨어의 역할

현재 신경망 VLSI와 하드웨어는 기존의 VLSI 기술인 실리콘 즉 CMOS 기술이 가장 널리 쓰이고 있다. 초기의 신경망 VLSI는 소규모 즉 1개의 처리소자를 가지며 연산고리는 시분할 방식으로 사용되는 것에 불과하였으나, 최근의 연구는 CMOS나 변형된 CMOS 공정으로 다위칩당 수천 연산고리의 용량에서 제한적이지만 수만 연산고리까지 구현하고 있으며 실리콘 웨이퍼를 통째로 1개의 칩처럼 이용하기에 이르렀다. 국내에서도 여러가지 종류의 신경망 VLSI를 연구하고 있으며 특히 현재까지 최대 규모의 용량이라 할 수 있는 십삼만 연산고리를 집적시킨 범용 하이브리드 신경망 VLSI인 URAN이 개발되었다[20].

신경망의 VLSI화 즉 하드웨어 구현은 여러가지 필요성과 문제점을 가지고 있으나, 그 주된 필요성은 기존의 컴퓨터 하드웨어상에서 시뮬레이션할 때 생기는 처리 속도와 규모의 제한을 하드웨어로 극복하고자 함에 있다.

일반적으로 신경망 하드웨어의 기술적 목표를 요약하면

대규모 용량의 신경망
동작의 고속성 즉 고속 병렬 처리
가변의 프로그램 즉 학습에 따른 연산고리 강도의 변경
경제성과 적절한 가격

을 들 수 있다. 이와 같은 목적의 소자는 LCD, GaAs 계열 반도체등의 광학적 구현도 생각할 수 있으나, Bipolar, CCD 특히 CMOS등의 기존의 VLSI 기술인 실리콘 기술이 보편화되어 있다. 이는 실리콘 VLSI 기술이 현재 가장 잘 발달된 기술이이서 실용화의 전제조건인 상대적으로 저렴한 가격과 신뢰도를 가능하게 하기 때문이다.

실리콘에 집적화시키는 신경망은 궁극적으로 방대한 규모의 연결고리(synapse)와 신경세포(neuron)가 연결된 망으로서, 연결고리는 곱셈과 덧셈 기능으로 모델링할 수 있으며 신경세포는 비선형 결정함수로 모델링 할 수 있다. 이와 같은 기능을 구현하는 데는 그 회로 설계 개념에 따라 디지탈과 아날로그로 구분할 수 있다. 여기에 추가하여 아날로그와 디지탈의 혼성인 하이브리드방식으로 설계할 수도 있다. 디지탈 방식과 아날로그 방식은 그 시스템이나 VLSI 제조공정에서 가지는 특성이 다르기 때문에 각각의 방식이 가지는 장, 단점도 서로 상반된다.

디지탈 방식으로 신경망 칩을 구현하면 우선 설계가 비교적 용이하다는 것을 제시할 수 있다. 그리고 디지탈인 관계로 임의의 정밀도를 구현할 수 있고 프로그램이나 취급이 용이하다는 점도 큰 장점 중의 하나이다. 그러나, 신경망 VLSI, 즉 하드웨어의 주 목적인 방대한 용량의 구현은 제한을 받는 데, 이는 연결고리의 기본 기능의 하나인 곱셈 연산에 상당한 칩 면적을 소비해야 하기 때문이다. 그리고 전체적인 동작을 완전한 비동기 동작으로 구현하기가 어려워 구조적인 문제점이 되기도 한다.

아날로그 방식으로 신경망 칩을 구현하면 가장 큰 장점으로 집적도가 높다는 점을 들 수 있다. 회로 내용에 따라 다르겠으나 비동기성 동작을 자연스럽게 얻을 수 있고 넓은 동작 범위를 가질 수도 있다. 그러나 아날로그 방식의 가장 큰 단점은 특수 제작 공정을 필요로 하여 구현이 어렵다는 점을 들 수 있다. 그리고 기술적으로 높은 임의의 정밀도를 얻기가 어렵거나 프로그램에 제한을 받기 쉽다.

아날로그와 디지탈의 하이브리드 방식은 여러가지 형태를 생각 할 수 있으나, 근본 동기는 디지탈 방식에서 얻을 수 있는 정밀도와 프로그램의 용이성을 얻으며 아날로그 방식의 장점인 소형 연결고리를 구현하고자 하는 것이다. 일반 디지탈 제작공정을 사용하며, 연결고리는 아날로그로 동작하고 정밀도나 사용과정은 디지탈에 준한다면 앞에서 언급한 디지탈의 장점과 아날로그의 장점만을 가질 수 있게 된다.

신경망 VLSI 성격으로 고려되는 다른 측면으로는 범용성 혹은 제한없는 확장성을 가지거나 혹은 특수한 용도 혹은 제한된 구조의 상태 여부에 따라 다른 장, 단점을 가지게 된다. 이는 아직 특정 구조나 학습방법 그리고 전기적인 신경 상태 변환 방식에 따른 일반화된 해결방안이 없기 때문에, 각각의 cpu 구조나 성능에 있어서는 다른 기준의 장, 단점을 가졌던 것 처럼 신경망 VLSI도 서로 다른 기준의 장, 단점을 가지므로 공통의 성능 기준으로는 연결고리 연산속도나 정밀도 및 용량을 단위로 하고 있다.

3.2 고속화 그리고 고집적화 과정의 신경망 VLSI

초기의 신경망 하드웨어는 개별 소자나 기존의 디지탈 로직, 연산증폭기를 이용하여 구성되었으며, 일부는 WIZARD와 같이 오래전에 상용화 되기도 하였다. 신경망 VLSI의 상용화 모델의 초기 단계는 Analog Neuro Processor(ANP)나 neuro bit slice를 들 수 있으며, 이들은 각각 1개의 뉴우론과 8개의 뉴우론을 가지고 있으며 연결고리의 등가적 구현은 한 순간에 하나씩 직렬 입력에 의한 시분할 방식으로 이루어졌다. 연결고리의 강도는 임의로 변경 가능하였고, ANP는 Bipolar 공정으로 제작된 특수한 곱셈형 D/A변환기와 시분할 샘플링을 이용한 아날로그 식이며 Micro Device사 제품은 MOS 공정으로 제작되었으며 AND 게이트를 이용한 1 비트씩 직렬 곱셈형 연결고리로 시분할적 wired-OR의 개념으로 동작한다. 이에 대하여 비교적 범용성, 확장성에서 어느 정도 개선된 stochastic pulse width로 구현하는 AND 게이트형 연결고리를 채택하여, wired-OR를 통한 신경망 VLSI가 최근에 출현했다. 비교적 확장성을 가지는 이 제품은 칩 세트 개념으로 사용되며 32×32 규모의 단위 칩으로 구성된다. 이 VLSI의 제한점으로는 디지탈 방식임에도 불구하고 비교적 정밀도가 낮고, 확장성을 가지나 그 규모가 stotchastic pulse operation에 따른 주기와 전체 크기에 의하여 제한을 가지는 점이다. 신경망 VLSI 그 자체로서 상품화 되어있는 것으로 ETANN 칩이 있으며, 상품화된 모델중 상위 성능과 가격을 가진다. 이는 아날로그 신경망 VLSI로서 특수한 CMOS 공정을 사용하여 nonvolatile 아날로그 연결강도 메모리를 내장하고 있다. 208핀 VLSI로 칩에서 외부와의 연결은 시분할 방식으로 이루어지고 있다. 디지탈 방식의 대단위 신경망 VLSI인 Adaptive Solution사의 CNAPS는 칩당 64개의 처리소자를 가진 무척 칩 크기가 큰 디지탈 방식으로 워크스테이션의 가속기 형태로 상용화되어 있다.

이와 같이 여러 형태의 신경망 VLSI가 상용화되어 있으나, 이 외에도 실험실에서 여러가지 방식으로 대규모 집적화된 신경망 VLSI가 연구되고 있다. 특수용도로서는 6인치 웨이퍼 스케일 집적화와 1 비트 연결고리 강도를 가지는 39,000 연결고리의 단일칩 집적화를 들 수 있다.

3.3 아날로그-디지탈 하이브리드 신경망 VLSI-URAN

아날로그나 디지탈로 신경망 칩을 구현하면 집적화 과정에서 여러 가지의 차이가 있으며 각각의 서로 상반된 장점과 단점, 즉 어느 한 쪽의 장점을 다른 면에서는 단점이 된다. 일반적으로 종래의 아날로그나 혼합방식에 의한 신경망 칩은 정밀도나 속도, 적응성에 있어 제한을 많이 받고 있다. 이와 같은 제한점들이 새로운 하이브리드 신경망 회로에서는 크게 개선되었는데, 기본적 특징은 연결 강도의 신축성을 강화시킨 MOSFET저항 원칩프로리를 이용한 곱셈연산에 있다. 그리고 원칩프로리 전류의 개폐를 제어하는 트랜지스터회에는 아날로그 단위 회로들이 제거된 짐도 큰 요인이다. URAN 신경망 칩은 다양한 연결 구조와 정밀도 개선이 가능한 그림 1의 92×92의 기본 모듈 16개로 이루어져, 임의의 외부제어를 통하여 재구성 가능한 특성의 구조이다.

URAN 칩은 대부분의 동작이 개념적으로는 아날로그 동작을 하며, 특히 디코더를 제외한 나머지 부분은 스위치와 정적 동작 특성을 가지는 MOSFET 저항 원칩프로리로 이루어졌다. 따라서 제작 기술이 발전함에 따른 용량과 속도를 개선할 수 있다. 신경망의 연결프로리는 원칩강도 저장 메모리를 포함하여 그림 2와 같이 9개의 트랜지스터로 이루어지며, 단위 구조는 원칩프로리간 배선의 여유까지 포함하여도 1.0 마이크론 표준 CMOS공정으로 900만평 마이크로미터 크기를 가진다. 기본적인 특성중 선형특성은 0.25%의 왜율을 가지고, 256계조 이상이 가능하다.

URAN의 원칩프로리 전자회로는 양극성을 가지는 입력에의 선형 전류원으로, 아날로그 샘플링식 신호처리와 같은 방법의 스위칭으로 원칩프로리 곱셈을 가능하게 한다. 원칩강도값의 전류치 변환은 MOSFET의 triode영역에서의 채널 저항을 선형화시키는 기술로 이루어졌으며 8 bit 이상의 정밀도를 가짐이 입증되었다. 원칩프로리의 정밀도 향상이나 구조의 대규모화 구현을 독립적인 양극성 전류원의 wired-OR 확장성으로 실현된다. 특히 기존의 디지털 방식에서 필수적인 클록이나 동기화 과정이 전혀 필요없다는 점이 확장성이나 변형성에 있어 큰 장점이다. URAN 신경

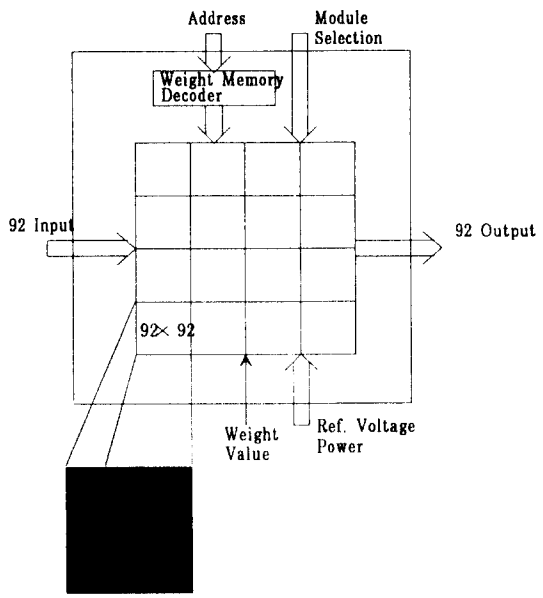

그림 1. URAN 칩 구조

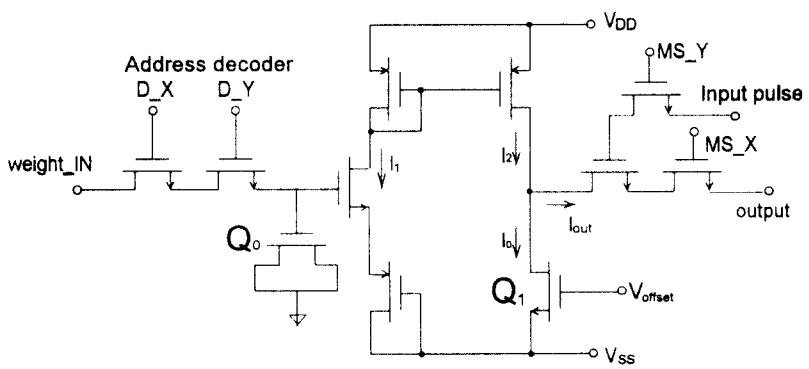

그림 2. 단위 Cell 구조

망 칩을 사용하는 측면에서는 전원과 함께 몇몇 기준 전압을 인가하는 이외에는 디지탈 신경망칩과 동일하게 사용할 수 있으며 타이밍 상의 제한이 거의 없이 완벽한 비동기성을 가진다. 칩의 구현은 신경 상태의 정의에 있어 단순한 디지탈 이외에 여러가지 펄스식 동작도 가능하므로 집적화의 특수성을 고려하여 연결고리와 처리소자의 칩 세트 개념으로 개발하고 있다. 처리 속도는 칩당 최대로는 340 Giga cps를 가지며 48개의 칩으로 PCB를 구성하면 16 Tera cps의 성능이 구현 가능하다.

Ⅳ. URAN 칩을 이용한 음성인식 시스템 설계

본절에서는 여러 형태의 URAN칩 중에서 가장 초기에 개발된 256개의 시냅스를 갖는 KTA11을 이용한 음성인식 시스템 설계에 관해 설명한다[21][22]. 보드를 설계한다. 이용된 신경망 알고리즘은 3층 구조를 갖는 MLP이다. 칩에 인가되는 음성 특징으로는 16개의 필터 뱅크 출력을 가정하였으며 연결강도값의 학습은 별도의 workstation상에서 off line으로 이루어진다. 시스템의 실시간 구현과 연결강도값 적재, 여러가지 제어 동작, 비선형 활성화 작용, 인식 결과의 결정 등을 위해서 DSP 보드를 이용한다. 신경망칩의 모든 인터페이스 회로는 FPGA(Field Programmable Gate Array)의 일종인 Xlinx를 이용해 설계하였다.

아날로그-디지탈 혼합형인 URAN은 기본 cell을 구성하는 아날로그 부분과 디코더 입력과 입력펄스를 가하는 디지탈 부분으로 크게 나눌 수 있다. 칩의 구조는 그림 3과 같이 8개의 입력단과 8개의 출력단이 있으며 4개의 모듈은 각각 64개의 cell로 구성되어 있다. 8개의 출력은 각각 32개의 cell이 wired-OR에 의해 연결되어 있으며, 모듈 선택에 의해 8개 단위로 cell을 선택할 수 있다.

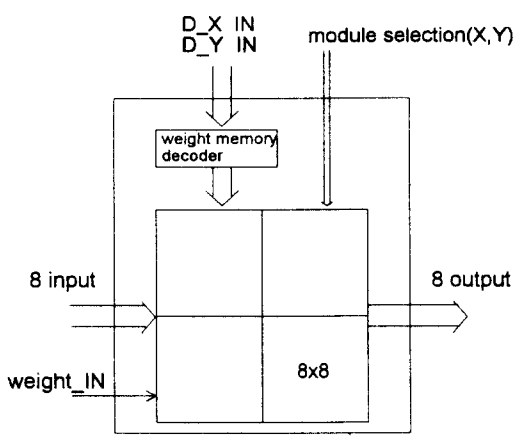

그림 3. KTA11의 구조

4.1 MLP 구조 및 학습과정

본 연구에 사용된 MLP 구조는 그림 4와 같은 MLP 구조를 갖고 있으며, 각 층간의 연결고리들은 fully connected 되어 있다. 입력층으로는 앞절의 전처리 과정에서 얻어진 16차 벡터의 음성 프레임이 순차적으로 인가되며, 칩의 출력이 8개인 특성을 고려하여 은닉층에는 16개의 node가 존재할 수 있도록 하였으며, 출력층에는 숫자음을 나타내는 node 11개로 이루어져 있다. 각 node들을 연결시켜주는 연결가중이들은 일성한 화자들의 데이타들로부터 학습되어진다.

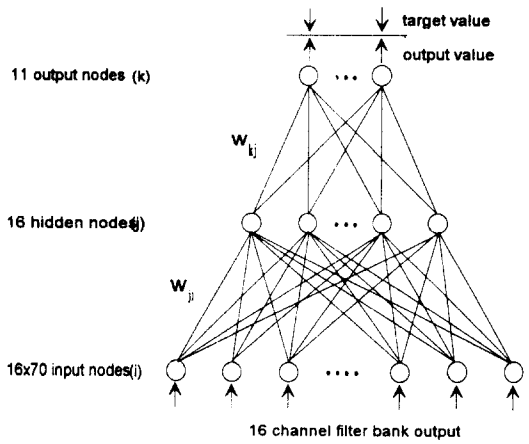

그림 4. MLP 구조

다층 퍼셉트론의 학습을 위한 오류 역전파 학습 알고리즘이 수행과정은 다음과 같다. 입력층의 각 뉴런에 입력 패턴이 가해지면 각 뉴런에서는 출력함수에 따라 활성화된 값이 은닉층에 전달되고 마지막에는 출력층에서 신호를 출력하게 된다. 이때 출력층의 k번째 뉴런에서 얻은 출력값 (O_k)과 k번째 뉴런에 대한 기대값($Target_k$)을 비교하여 다음 식에 의해 정의되는 신경망의 total sum-of-squared-error를 계산한다.

$$E = \frac{1}{2} \sum_{\text{all patterns}} \sum_{k} (Target_k - O_k)^2 \quad (1)$$

(1) 식에 의해 정의된 error를 줄여 나가는 방향으로 연결강도를 조절하고, 상위층에서 역전파하여 하위층에서는 이를 근거로 다시 자기층의 연결강도를 조정해나간다. 일반적으로 다층 퍼셉트론에서는 신경망의 total sum of squared error가 임계치 이하로 수렴될 때가 학습된 상태이다. 각 뉴런의 출력함수에 의한 활성화값은 다음과 같이 정의 된다.

$$O_k = \frac{1}{1 + e^{-\sum \omega_{kj} O_j}} \quad (2)$$

단, O_j는 은닉층 뉴런 j의 활성화 값이고, ω_{kj}는 은닉층 뉴런 j에서 출력층 뉴런 k로의 연결강도를 표시한다. 은닉층 뉴런의 활성화 값 또한 입력층 뉴런의 활성화 값과 입력층 뉴런에서 은닉층 뉴런으로의 연결강도에 의해 결정된다. 연결강도의 변경은 아래식에 따라 이루어진다.

$$\omega_{kj}(t) = \eta \Delta \omega_{kj} + \gamma \omega_{kj}(t-1) \quad (3)$$

단, Δ는 학습률이고, γ는 학습속도를 빠르게 하기 위한 관성률, t는 학습 횟수이다. 학습률 Δ는 0.1, 관성률 γ는 0.9로 하였으며 연결강도의 초기값은 -0.5에서 0.5 사이의 랜덤값으로 주어진다. 학습과정은 total-sum-of-squared-error가 0.01에 도달하거나 제한된 학습 횟수에 도달할 때까지 반복된다.

4.2 전체 시스템의 구조 및 동작

전체 시스템의 개략적인 구성은 그림 5와 같이 크게 PC와 DSP 보드 및 신경망 보드로 구성된다.

1) PC

먼저 PC에서는 신경망 보드에서 필요한 2개의 FPGA 칩에 대한 정보와 학습된 연결강도값, DSP Program 등을 보유하며 User Interface를 담당한다. 또한 신경망 보드의 메모리에 연결강도값 등을 적재하며 DSP 보드에는 음성인식 main 프로그램을 적재한다. 이후 PC는 실제 인식 과정에서는 별다른 일을 하지 않고 있다가 신경망 보드에서 음성인식 과정이 종료되면 그 결과를 읽어 들여 화면상으로 출력하게 된다.

2) DSP 보드

DSP 보드에서는 마이크로부터 입력되는 음성신호의 끝점을 검출하고 음성의 특징을 추출해서 신경망 보드에 가해주는 역할을 한다. 입력음성이 끝날 때까지 매 프레임마다 10 msec의 간격으로 16차의 특징 벡터를 구하며, 이 값은 제어 신호와 함께 신경망 보드로 넘겨진다. 또한 신경망의 각 은닉층에서의 출력을 읽어서 활성화 작용을 한다. URAN은 활성화 작용을 하는 뉴런 구조가 없으므로 현재로는 DSP에서 활성화 작용을 하게 된다. 이렇게 계산된 활성화 값을 다시 신경망 보드에 가해 주어서 다음 계층의 입력으로 이용하게 한다. 이때 최종 출력층의 경우 출력값을 이용해서 인식단어를 결정하여 PC측에 전송한다.

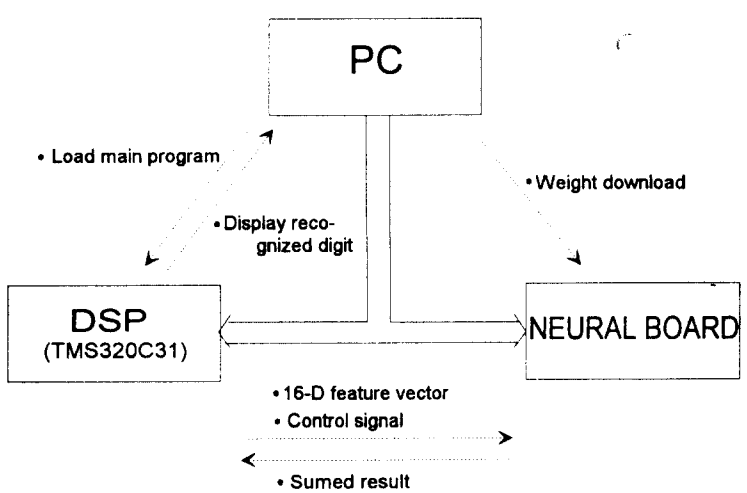

그림 5. 전체 시스템의 구조

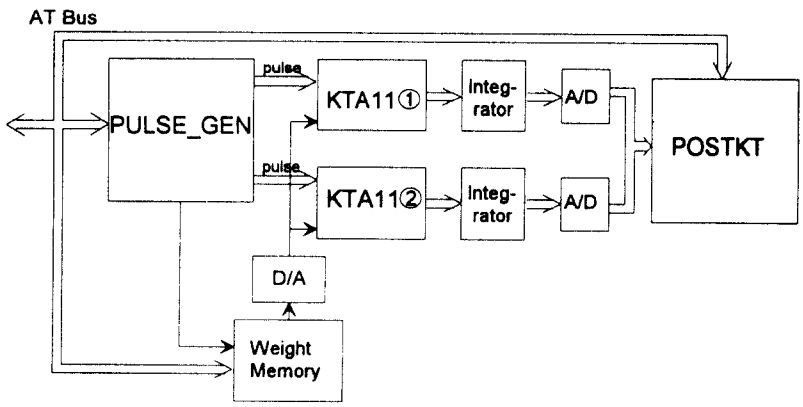

그림 6. URAN을 이용한 신경망 보드의 구조

3) 신경망 보드

신경망 보드는 입력 특징벡터를 받아서 매 프레임마다 곱셈과 합을 구하는 과정과 프레임별 결과를 축적하는 역할을 한다. 먼저 PC로부터 연결강도값이 적재된 후 DSP 보드로부터 인가되는 음성 특징값들을 입력으로한 신경망의 출력을 구한다. 이때 모든 음성 프레임에 대해서 출력값을 누적하며 음성 입력이 끝나면 누적된 출력값을 DSP보드로 전송한 뒤, 다음 계층의 입력을 기다리게 된다. 마지막 계층의출력을 구한 뒤 그 값을 DSP 보드에 전송하여 해당 인식단어를 결정하게 된다. 그림 6은 URAN중 가장 간단한 KTA11을 이용한 신경망 음성인식 보드의 구조를 나타낸다. 이때 PULSE_GEN과 POSTKT는 신경망 보드의 제어 신호를 발생시키기 위해서 FPGA로 설계된 2개의 제어칩에 붙혀진 이름이다.

시스템이 동작을 하면서 가장 먼저 하게 되는 것 중의 하나가 off-chip으로 학습된 연결강도값들을 연결가중값 메모리에 적재하는 것인데, 연결강도값들은 매프레임의 입력을 처리하기 전에 PC로부터 적재되어서 보드상의 3개의 32 kByte RAM에 저장된다. 연결강도값 적재를 위해 3개의 port를 이용해서 보드 RAM의 어드레스 디코딩을 하게 되는데, 이 부분에서는 연결강도를 적재하는 것 뿐 아니라 복잡하게 되어 있는 KTA11의 디코더 구조 때문에 소요되는 클럭수를 최소로하면서 모든 cell들을 선택하고 연결강도값을 적재하기 위한 최적의 시이퀀스 또한 적재하게 된다. PC상에서 디코더 입력과 연결가중치의 동기를 맞추어 시이퀀스를 발생시켜 화일로 만들어 놓은 다음 AT Bus를 통해 메모리에 적재하게 된다. 두개의 URAN 칩을 사용했기 때문에 16개의 입력과 16개의 출력을 얻을 수 있고, 칩 출력 값을 축적할 수 있는 적분기와 매 프레임이 끝날 때마다 축적된 값을 디지탈로 변환할 수 있도록 A/D 변환기가 각 출력단에 연결되어 있다.

신경망 보드의 제어 신호는 2개의 FPGA로 설계된 PULSE_GEN과 POSKT에 의해서 발생된다. PULSE_GEN은 AT Bus로부터 들어오는 어드레스를 디코딩해서 연결강도값을 메모리에 적재하는 작용과 입력 특징 벡터를 받아서 그에 상응하는 펄스를 만들어 두개의 URAN 칩 각각에 가해주는 역할을 하게 된다. POSTKT는 매 프레임마다 PULSE_GEN으로부터 제어 신호를 받아 URAN 출력단에서의 A/D 된 결과를 저장하고 축적하는 동작을 하게 된다. 일단 DSP 보드로부터 제어 포트를 통해 입력 프레임 동기 신호를 받으면 이때부터 PULSE_GEN의 어드레스 카운터에 의해 메모리로부터 디코더 시이퀀스를 읽어 칩내의 cell을 선택하면서 동시에 연결강도값을 읽어 D/A 변환기를 통해 두개의 칩에 적재하게 되는데, 매 프레임 끝마다 적재가 끝났음을 알리는 신호를 받아 정상 동작에 들어가게 된다. URAN은 연결강도값에 비례하는 전류를 PULSE_GEN에서 인가하는 펄스의 갯수만큼 흘려줌으로써 곱셈작용을 하게 된다. 이러한 전류는 URAN외부에 별도로 부착되는 커패시터에 축적되며 한 프레임이 끝날때마다 A/D변환되어 POSTKT에서 축적된다.

FPGA 시뮬레이션에 의한 검증 결과 한프레임에

대한 연산시 신경망 보드에서 소요되는 클럭 수는 연결가중치 적재시 289 클럭, 뉴런 변환시 512클럭 등 약 800클럭이 소요됨을 알 수 있었다. 따라서 1 MHz의 클럭을 사용할 경우 1 ms에 연산이 끝나게 된다. 이 클럭수는 인식 대상 낱어의 수가 늘어나도 복잡이 적용되는 것으로 낱지 사용되는 칩의 갯수 및 회로의 크기만 증가시켜 주면 같은 인식 시간이 소요되는 시스템의 구현이 가능하다.

V. 낮은 정밀도 계산에 의한 시뮬레이션

본 절에서는 낮은 정밀도의 입출력 및 연결강도를 사용한 신경망의 성능 및 sigmoid 출력 함수로 학습된 신경망에 대해 선형 출력함수를 사용하여 테스트 한 경우의 성능을 검토하였으며, 구간 선형 출력함수를 이용한 신경망의 학습을 시도하였다. 또한 잡음 환경하에서의 신경망의 인식 성능을 분석하였다.

5.1 URAN 칩을 위한 고려사항

다층 퍼셉트론을 학습하기 위한 오류 역전파 알고리즘을 수행하는데 있어서 가장 큰 제약사항은 높은 정밀도 계산이 필요하다는 점이다. 오류 역전파 알고리즘을 병렬처리 구조의 신경망 하드웨어 상에서 효율적으로 구현하기 위해서는 다음의 세 가지 사항이 고려되어야 한다. 즉 오류를 역전파하기 위한 연결강도의 정밀도, 오류계산과정의 정밀도 및 scaling, 그리고 출력함수의 구현등이다[4][23][24].

본 시뮬레이션에서는 입출력의 정밀도 및 연결강도의 정밀도 그리고 출력함수 등이 고려되었다. URAN 칩의 연결강도의 정밀도는 최대 8 bit이므로 우선 8 bit 이하의 정밀도를 가지는 연결강도의 유용성을 검토하기 위해 일반적인 학습과정을 통해 얻어진 부동소숫점 연결강도의 정밀도를 낮추어 테스트해 보도록 한다. 또한 정밀도를 낮춘 특징 벡터와 함께 가장 간단화된 binary 특징벡터를 입력으로하여 신경망 테스트를 수행한다.

두번째로, sigmoid 비선형 출력함수에 의해 학습된 연결강도값을 가진 선형 출력함수 신경망에 대한 테스트를 수행하며, 마지막으로 sigmoid 함수를 구간화해 얻은 구간 선형 출력함수를 이용하여 신경망의 학습 및 테스트를 수행한다. 이것은 비교적 간단한 선형 출력함수를 가지는 neuron 칩의 하드웨어 구현 가능성을 검토하기 위한 것이다.

5.2 음성 데이터 베이스 및 전처리 과정

신경망 인식 실험은 10개의 한국어 숫자음을 사용하였다. 남/여 각각 10명, 총 20명의 화자가 10번씩 발음한 2,000개의 데이터 중에서 남·여 각각 5명이 5번씩 발음한(5회×10 숫자음×10명 = 500개) 데이터를 학습하는데 사용하였다. 학습에 참가한 화자가 발음한 나머지 5번씩의 음성 데이타 500개는 다화자 종속 인식 실험에, 학습에 참가하지 않은 남녀 각각 5명이 10번씩 발음한 1,000개의 음성 데이터를 화자독립 인식실험에 사용하였다.

음성신호는 조용한 사무실 환경에서 탁상용 마이크로 녹음되었으며 차단 주파수가 4.7 kHz인 아나로그 저역통과필터로 여과된 뒤, 10 kHz 샘플링 주파수, 12 bit로 A/D 변환되었다. 잡음 환경을 모의 실험하기 위해서 컴퓨터에 의해 Gaussian 분포를 가지는 백색잡음을 생성하여 신호 대 잡음비가 각각 30 dB, 20 dB, 10 dB, 0 dB이 되도록 위래의 음성 신호와 잡음을 섞은 뒤 다시 4.7 kHz의 차단 주파수를 가지는 디지탈 필터를 통과시켜 잡음이 섞인 음성 신호를 만들었다. 음성부분만 검출된 숫자음 데이터를 특징벡터를 평탄화시키기 위하여, 0.95의 비율로 preemphasis를 취한 후 20 msec의 Hamming window를 10 msec씩 이동시키며 얻은 각 프레임마다 512 포인트 FFT를 수행하였다. 이로부터 17 channel의 critical-band 필터 뱅크의 각 대역 에너지를 평균하여 매 프레임마다 17차의 부동소수점 특징벡터를 구하였다. 실제 신경망에 사용된 입력 벡터는 -1과 1 사이의 값으로 정규화 되었다. 또한 낮은 정밀도의 입력을 검토하기 위해서 부동 소수점 데이터는 소수 4자리, 2자리 및 1자리로 제한되어 사용되었다.

URAN 칩은 각 뉴런 별로 디지탈 펄스열의 입력 데이터를 받아들이게 되어 있으므로 binary 입력이 가장 효율적이다. 본 연구에서는 각 프레임 별로 17 bit로 구성된 이진 스펙트럼을 특징벡터로 사용하여 그 성능을 분석하였다. 이진 스펙트럼을 구하는 과정은 다음과 같다. 먼저 음성 신호 중에서 음운정보는 주로 스펙트럼의 peak부분에 분포되어 있다는 가정하에 LPC 스펙트럼의 2차 미분값이 임계치를 넘는 경우 "1"상태로 클램핑 시킨다. 다음에 주파수 대역을 17개의 critical-band로 구성하여 "1"로 클램핑된 주파수 성분이 있는 대역을 다시 "1"로 클램핑 시킨다. 이 과정을 통해 매 프레임당 17 bit로 표현된 특징벡터가 구해진다[25].

5.3 잡음환경에서의 화자독립 숫자음 인식 실험

잡음 환경에서의 인식 수행에 앞서 적절한 은닉층 노드 갯수를 결정하기 위해서 20개에서 50개까지 10개 단위로 은닉층 노드 갯수를 변경시키면서 인식률을 살펴 보았으며 그 결과는 표 1과 같다. 실험결과 은닉층 노드 갯수에 따른 성능의 차이가 그렇게 크지 않고 그중 은닉층 노드가 30개일 때의 성능이 비교적 좋았으므로 앞으로의 실험에서의 은닉층 노드는 모두 30개로 고정하였다.

표 1. 은닉층 노드 갯수에 따른 인식률

# of Hidden Units	floating point input			binary input
	4 decimal	2 decimal	1 decimal	1 bit
20	96.5 %	96.2 %	96.1 %	89.3 %
30	97.6 %	97.5 %	97.5 %	90.8 %
40	96.6 %	96.6 %	96.4 %	91.0 %
50	96.8 %	96.9 %	97.3 %	88.9 %

첫번째 수행한 실험은 sigmoid 비선형 출력함수를 사용한 신경망에 대해 입력과 출력, 연결강도의 정밀도를 각각 소수점 4자리, 2자리, 1자리 등으로 변경하여 학습 및 테스트를 수행하였다. 각각의 다층 퍼셉트론의 은닉층 뉴런 수는 30개, 학습 횟수는 소수점 2자리 및 1자리의 정밀도를 가지는 신경망의 경우 학습 횟수를 200 번으로 제한하였으며, 이진 입력을 가지는 신경망은 500 번으로 제한하였다. 모든 경우에서 다층 퍼셉트론의 출력은 소수점 2자리의 정밀도를 가진다. 입력의 정밀도가 소수점 2자리인 경우 연결강도의 정밀도도 소수점 2자리를 가지며, 입력의 정밀도가 소수점 1자리인 경우 연결강도의 정밀도도 소수점 한자리를 갖도록 하였다. 이진 스펙트럼 입력의 경우 연결강도의 정밀도는 소수점 2자리를 갖도록 하였다. 화자독립으로 숫자음을 인식한 결과를 보면 표 2와 같이 정밀도에 따른 차이는 거의 없었다.

표 2. Sigmoid 출력함수 신경망의 인식 결과

SNR ratio	floating point input			binary input
	4 decimal	2 decimal	1 decimal	1 bit
clean	97.3 %	97.1 %	97.2 %	90.8 %
30 dB	96.2 %	96.2 %	96.7 %	90.9 %
20 dB	88.9 %	88.8 %	89.8 %	87.5 %
10 dB	60.2 %	60.4 %	60.3 %	67.2 %
0 dB	29.5 %	29.8 %	29.7 %	38.4 %

두번째로 수행한 실험은 선형 출력함수의 테스트이다. 선형 출력함수를 사용한 경우 학습이 이루어지지 않으므로 sigmoid 함수를 사용하여 학습을 수행시킨 뒤 다음 식으로 정의된 선형 출력함수를 가지는 신경망으로 테스트하였다.

$$y = a(\sum_k \omega_{kj} O_j) + b, \qquad -A \le \sum_k \omega_{kj} O_j \le A \qquad (4)$$

식 (4)로 주어진 선형 출력함수의 정의 영역이 sum-of-weighted-input 값의 범위를 포함할 경우 인식률의 변화는 거의 없었다. 표 3에 a = 0.1, b = 0.5, A = 5인 경우에 대한 화자독립 숫자음 인식률을 비교하였다.

표 3. 선형 출력함수 신경망의 인식 결과

SNR ratio	floating point input			binary input
	4 decimal	2 decimal	1 decimal	1 bit
clean	97.7 %	97.5 %	97.2 %	90.7 %
30 dB	96.4 %	96.2 %	96.6 %	90.5 %
20 dB	90.2 %	90.1 %	91.3 %	86.6 %
10 dB	59.8 %	59.8 %	59.9 %	68.0 %
0 dB	30.0 %	30.8 %	29.5 %	38.5 %

이상의 실험에서 학습이 이루어진 오류 역전파 알고리즘은 테스트 과정에서 연결강도, 입력 및 출력의 정밀도의 영향을 받지 않음을 알 수 있다. 또한 입력 및 출력이 소수점 2자리 및 1자리의 정밀도를 가진 경우에도 학습이 이루어짐을 알 수 있다. 이는 각각 8 bit 및 4 bit 정도의 정밀도에 해당하는 것이다.

세번째로 수행한 실험은 선형 출력함수를 이용한 학습 가능성이다. 식 (4)와 같이 단순한 선형 출력함수를 사용한 신경망의 경우 학습이 전혀 이루어지지 않아서 본 연구에서는 sigmoid 비선형 함수를 근사화한 구간선형함수를 유도하여 학습을 시도하였다. 즉, sigmoid 함수의 정의 영역을 세구간으로 나눈 뒤 sigmoid 함수와 구간선형함수 사이의 mean-squared-error (MSE)가 최소가 되도록 하여 식 (5)와 같은 구간선형함수를 유도하였다. 소수점 3자리까지 계산했을 때 sigmoid 함수와 식 (5)사이의 MSE는 0이 된다.

$$\begin{aligned} y &= 0.0, && x < -7.6 \\ y &= 0.0087x + 0.066, && -7.7 \le x < -2.2 \\ y &= 0.206x + 0.5, && -2.2 \le x < 2.2 \\ y &= 0.087x + 0.934, && 2.2 \le x < 7.6 \\ y &= 1.0, && x > 7.6 \end{aligned} \qquad (5)$$

식 (5)와 같은 구선형 출력함수를 사용한 신경망에 대한 학습 시간은 sigmoid 비선형 출력함수를 사용한 경우와 거의 비슷한(Sun Sparc 10을 사용하여 약 1일)시간이 걸렸으며, 그때의 인식결과는 표 4에 나타나 있다.

표 4. 구간선형 출력함수 신경망의 인식 결과

SNR ratio	floating point input			binary input
	4 decimal	2 decimal	1 decimal	1 bit
clean	97.2 %	97.1 %	96.6 %	90.1 %
30 dB	96.4 %	96.4 %	95.9 %	89.9 %
20 dB	91.3 %	91.3 %	90.9 %	86.4 %
10 dB	61.5 %	61.6 %	58.8 %	65.2 %
0 dB	32.5 %	32.6 %	34.6 %	37.7 %

이때, 각 신경망은 30개의 은닉층 뉴런을 가지도록 했으며, 학습 횟수는 앞에서의 실험과 같은 횟수로 제한되었다. 표 4에서 보는 바와 같이 출력의 정밀도에 따른 성능 변화는 거의 없었다. 이것은 신경망의 성능이 구간선형 출력함수의 기울기에 민감하지 않음을 나타내며, 선형 출력함수의 기울기가 성능에 큰 영향을 주지 않는 이유와 같은 맥락으로 볼 수 있다.

마지막으로 학습패턴에 백색잡음을 첨가하여 인식실험을 수행하였다. 이는 잡음의 영향을 평가하기 위한 것으로 표 5와 같이 잡음이 섞이지 않은 경우 학습패턴에 대한 인식률은 100%이나 SNR이 낮아질수록 인식률은 급격하게 저하된다.

표 5. 백색잡음이 섞인 학습패턴에 대한 인식 결과

SNR ratio	sigmoidal units		piecewise linear units	
	4 decimal	2 decimal	4 decimal	1 bit
clean	100 %	100 %	100 %	100 %
30 dB	100 %	99.6 %	100 %	100 %
20 dB	94.0 %	96.2 %	94.8 %	97.8 %
10 dB	68.0 %	78.8 %	70.4 %	76.4 %
0 dB	37.0 %	43.2 %	40.2 %	42.6 %

이상의 실험을 통해 Host에서 학습된 연결강도 값을 URAN 칩에 적재하여 동작시킬 경우 낮은 정밀도의 입출력 및 연결강도, 선형 출력함수를 가지는 뉴런을 사용하여도 성능 저하없이 음성인식을 수행할 수 있으며 sigmoid 함수를 근사화한 구간선형 출력함수를 사용할 경우 sigmoid 함수를 사용한 경우와 거의 같은 정도로 학습이 이루어지며 테스트시에도 성능 저하가 거의 없음을 알 수 있었다. 더우기 잡음 환경인 경우에서도 입출력이나 연결강도의 정밀도가 낮아질 때 신경망의 성능은 저하되지 않았다. 따라서 뉴런칩의 출력함수를 구간 선형함수로 구성한다면 URAN 칩을 이용한 on-chip 학습도 가능할 것이다. 또한 이진 입력을 사용할 경우 SNR이 10 dB 이하인 상황에서 오히려 더 인식률이 높다는 사실을 알 수 있다.

VI. 결 론

본 글에서는 음성인식에 적용된 신경망 구조를 알아보았다. 크게 정적 신경망과 동적 신경망으로 분류한 뒤 여러가지 음성인식 신경망에 대해서 살펴보았다. 또한 신경망 VLSI와 국내에서 개발된 신경망 VLSI인 URAN에 대해서 알아보았으며, URAN을 이용한 음성인식 시스템을 설계하고 아날로그 디지탈 하이브리드 신경망 칩의 하드웨어 인터페이스를 개발할 수 있는 FPGA 환경을 구축하였다. 시뮬레이션을 통해 낮은 정밀도의 입출력 및 연결강도, 선형 출력함수를 가지는 뉴런을 사용하여 성능 저하없이 음성인식을 수행할 수 있었으며, 구간선형 출력함수를 사용할 경우 sigmoid 함수를 사용한 경우와 거의 같은 정도로 학습과 인식이 이루어짐을 알 수 있었다. 또한 잡음 환경에서도 낮은 정밀도를 사용한 신경망의 성능저하가 크지 않음을 확인하였다.

만족할 만한 성능의 신경망 음성인식 시스템을 구현하기 위해서는 신경망의 강점인 학습 가능과 대단위 병렬 동작성의 특성을 충분히 살릴 수 있고 또한 신경망의 약점인 시간 정렬과 계층적 시스템 구성의 난이성 등을 극복할 수 있는 신경망 구조가 요구된다. 또한 아날로그 또는 아날로그/디지탈 혼합 방식의 신경망 VLSI 칩 설계시 주변 칩과의 디지탈 인터페이스를 충분히 고려하면 좀더 효율적인 시스템 구성이 가능할 것이다. 일반적으로 음성인식을 위해서는 음성신호의 특징 자체에 대한 연구뿐 아니라 문장의 문법과 의미에 대한 연구가 필수적이다. 따라서 이러한 문법과 의미를 처리할 수 있는 구조가 신경망 음성인식 시스템에 부가되어야 할 것이다.

참 고 문 헌

1. A. Waibel, "Neural Network Approaches for Speech Recognition," *Advances in Speech Signal Processing*, Marcel Dekker, INC., pp. 557-590, 1992.

2. 이수영, "신경망을 이용한 음성 인식," 한국통신학회지 제 9권 제 11호, pp. 18-23, 1992년 2월.
3. H. Bourlard, N. Morgan and S. Renals, "Neural Nets and Hidden Markov Models : Review and Generalizations," *Speech Communication*, vol. 11, no. 2-3, pp. 237-246, June 1992.
4. R.M. Debenham and S.C.J. Garth, "Investigations into the Effect of Numerical Resolution on the Performance of Back Propagation," *Neural Networks from Models to Applications*, L. Personnaz and G. Dreyfus, Eds. I.D.S.E.T. Paris, pp. 725-755, 1989.
5. H. Bourlard and C. J. Wllekens, "Speech Pattern Discriminatin and Multilayer Perceptrons," *Computer Speech & Lanugage*, vol. 3, no. 1, pp. 1-19, Jan. 1989.
6. R. P., Lippman, and B. Gold, "Neural-net Classifiers Useful for Speech Recognition," *IEEE Int. Conf. Neural Networks*, vol. 4, June 1987, pp. 417-425.
7. M. Niranjan, F. Fallside, "Neural Networks and Radial Basis Functions Classifying Static Speech Patterns," *Computer Speech & Language*, vol. 4, no. 3, pp. 275-289, July 1990.
8. T. G. Kohonen, *Self-Organization and Associative Memory. 2nd Edition*, Springer-Verlag, Berlin, pp. 119-157, 1987.
9. T. G. Kohonen, G. Barna and R. Chrisely, "Statistical Pattern Recognition with Neural Networks : Benchmarking Studies," *IEEE, Proc. of ICNN*, vol. I, July 1988, pp. 61-68.
10. A. Waibel, "Modular Construction of Time-delay neural networks for speech recognition," *Neural Computation*, vol. 1, pp. 39-46, 1988.
11. P. Haffner, "Connectionist Word-Level Classification in Speech Recognition," *IEEE, Proc. ICASSP*, 1992, pp. 621-624.
12. H. Hild and A. Waibel, 1993, "Multi-Speaker/Speaker-Independent Architectures for the Multi-State Time Delay Neural Network," *IEEE, Proc. ICASSP*, vol. 1, 1993.
13. R. L. Watrous, L. Shastri, and A. H. Waibel, "Learned phonetic discrimination using connectionist networks," *Proc. European Conf. Speech Tech.*, Edinburgh, Sep. 1987, pp. 377-380.
14. J. Tebelskis, A. Waibel, B. Petek, et. al., "Continuous Speech Recognition by Linked Predictive Neural Networks," *Advances in Neural Information Processing Systems*, vol. 3, 1991, pp. 199-205.
15. E. Levin, "Speech recognition using hidden control neural network architecture," *IEEE, Proc. ICASSP*, vol. 1, Apr. 1990, pp. 433-436.
16. N. Morgan and H. Bourlard, "Continuous speech recognition using multi-layer perceptrons with hidden Markov models," *IEEE, Proc. ICASSP*, vol. 1, 1990, pp. 413-416.
17. H. Bourlard, C. J. Wellekens, "Links Between Markov Models and Multilayer Perceptrons," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 12, pp. 1167-1178, Dec. 1990.
18. T. Robinson and F. Fallside, "A recurrent error propagation network speech recognition system," *Computer Speech & Language*, vol. 5, no. 3, pp. 259-263, July 1991.
19. A. Mellouk, P. Gallinari, F. Rauscher, "Prediction and Discrimination in Neural Networks for Continuous Speech Recognition," *EUROSPEECH '93*, vol. 3, Sep. 1993, pp. 1603-1606.
20. 한일송, "신경망 VLSI 기술의 발달과 현재," 한국통신공학회지, 제 9권 제 11호, pp. 47-52, 1992년 11월.
21. Ki-Chul Kim, Il-Song Han, Jun-Hee Lee, and Hwang-Soo Lee, "Speaker Independent Digit Recognition with Reduced Representations for Neural Network VLSI Chip," *Proc. of World Congress on Neural Networks*, vol. 4, San Diego, June 1994, pp. 568-573.
22. 이준희, "URAN 신경망 칩을 이용한 숫자음 인식 시스템의 구현에 관한 연구," 한국과학기술원 정보 및 통신공학과 석사논문, 1994.
23. H. McCarter, "Back Propagation Implementation on the Adaptive Solutions CNAPS Neurocomputer Chip," *Advances in Neural Information Processing Systems 3*, R.P. Lippman, J.E. Moddy, and D.S. Touretzky, Eds. Dan Mateo, CA : Morgan Kaufman, pp. 1028-1030, 1991.
24. J.L. Holt and T.E. Baker, "Back Propagation Simulations Using Limited Precision Calculations," *Proc. of IJCNN'91*, vol. II, 1991, pp. 121-126.
25. K. C. Kim and J. W. Cho, "Robust Speech Recognition using Frequency Weighted All-Pole Model Sp-

ectrum," *Computer Processing of Chinese & Oriental Language*, vol. 5, no. 3 & 4, pp. 203-216, Nov. 1991.

석 용 호

- 1969년 10월 20일생
- 1992년 2월 : 한양대학교 전자공학과 졸업(공학사)
- 1994년 2월 : 한국과학기술원 전기 및 전자공학과 졸업(공학석사)
- 1994년 3월 ~ 현재 : 한국과학기술원 시울분원 정보 및 통신공학과 박사과정 재학중
- 주관심분야 : 음성인식, 신경망, 통계적 신호처리

김 기 철

- 1958년 4월 12일생
- 1983년 2월 : 한양대학교 전자공학과(공학사)
- 1985년 2월 : 한국과학기술원 전산학과(공학석사)
- 1992년 2월 : 한국과학기술원 전산학과(공학박사)
- 1992년 3월 ~ 1992년 7월 : 컴퓨터 응용기술(주) 개발부 차장
- 1992년 8월 ~ 현재 : 한국과학기술원 시울분원 정보 및 통신공학과 전임연구원
- 주관심분야 : 음성인식, 자연어처리, 신경망, 컴퓨터 구조

한 일 송

- 1956년 2월 1일생
- 1979년 2월 : 서울대학교 전자공학과(공학사)
- 1981년 2월 : 한국과학기술원 전기 및 전자공학과(공학석사)
- 1984년 2월 : 한국과학기술원 전기 및 전자공학과(공학박사)
- 1984년 5월 ~ 1985년 2월 : 한국과학기술원 전자공학부 연수연구원
- 1989년 4월 ~ 1990년 3월 : 영국 Edingurgh대학교 전기과 Visiting Academic
- 1985년 ~ 현재 : 한국전기통신공사 연구개발원 책임연구원(신경망 연구팀장)
- 주관심분야 : 신경망칩 기술과 응용시스템 개발

이 황 수

- 1952년 9월 19일생
- 1975년 2월 : 서울대학교 전기공학과(공학사)
- 1978년 8월 : 한국과학기술원 전기 및 전자공학과(공학석사)
- 1983년 2월 : 한국과학기술원 전기 및 전자공학과(공학박사)
- 1975년 1월 ~ 1975년 10월 : 현대조선중공업(주) 설계부 사원
- 1983년 3월 ~ 1989년 2월 : 한국과학기술원 전기 및 전자공학과 조교수
- 1989년 3월 ~ 1992년 1월 : 한국과학기술원 전기 및 전자공학과 부교수
- 1984년 4월 ~ 1985년 5월 : 미국 Stanford 대학교 Information Systems Lab. Post Doc. 연구원
- 1992년 2월 ~ 1993년 8월 : 한국과학기술원 서울분원 정보 및 통신공학과 부교수
- 1993년 9월 ~ 현재 : 한국과학기술원 서울분원 정보 및 통신공학과 교수
- 주관심분야 : 디지탈 통신, 이동통신, 신호처리(통신, 음성, 레이다)