신경망 VLSI 칩을 이용한 음성인식 시스템 설계 및 인식 실험

석 용 호* °, 김 기 철*, 한 일 송**, 이 황 수*
* : 한국과학기술원 정보및통신공학과 ** : 한국통신 연구개발원

A Speech Recognition System Design using Neural Network VLSI Chip and Recognition Experiments

Yong-Ho Suk* °, Ki-Chul Kim*, Il-Song Han**, and Hwan-Soo Lee*
* : Dep. of Information and Communication Eng. KAIST ** : Korea Telecom Research Center

요 약

본 논문에서는 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. 서론

신경망을 이용한 음성인식 방식의 장점은 입력을 이용해서 스스로 학습할 수 있으며 음성 신호에 내재된 특징을 자연스럽게 추출할 수 있다는 점이다. 또한 간단한 단위 구조의 반복으로 전체 시스템을 구현할 수 있으며 내부 결손과 잡음에 강인하다는 점이다 [1] [2] [3]. 음성인식에 이용되는 신경망은 입력 음성신호 또는 특징벡터의 시간축상의 변이 (time shift)에 대한 고려 방식에 따라 정적 신경망과 동적 신경망으로 나뉘게 된다. 본 논문에서는 정적 신경망의 일종인 다층 퍼셉트론 (Multi Layer Perceptron, MLP)을 이용하였다.

본 논문에서는 아날로그/디지탈 혼합방식 신경망 칩인 Universally Reconstructable Artificial Neural Network (URAN)을 이용한 실시간 음성인식 시스템을 설계하고, 이의 구현시 고려되어야 할 낮은 정밀도 계산을 통해 역전파 신경망의 성능을 검토한다.

제 II절에서는 신경망의 VLSI 구현에 대해서 알아보며 아날로그/디지탈 혼합 방식의 신경망 VLSI인 URAN에 대해 설명한다. 제 III절에서는 URAN 및 이를 이용한 음성인식 시스템의 설계에 대해서 기술한다. 제 IV절에서는 낮은 정밀도 계산에 의한 시뮬레이션을 통해 잡음이 존재하는 환경하에서의 신경망의 숫자음 인식 성능을 실험하고 그 결과를 분석하였다. 제 V절에서는 결론과 신경망을 이용한 음성인식 시스템의 성능향상에 필요한 연구 방향을 고찰하였다.

II. 신경망의 VLSI 구현

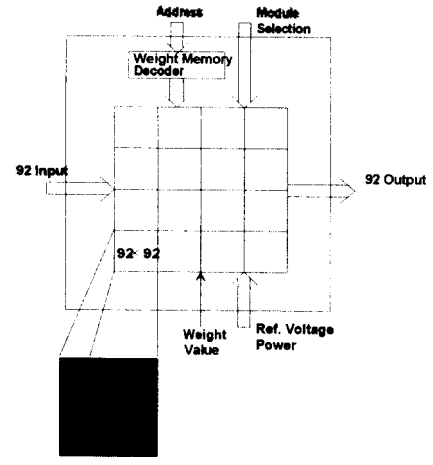
2.1 신경망 VLSI와 하드웨어의 역할

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 [4].

신경망 VLSI 구현 방법에는 임의의 정밀도를 구현할 수 있고 칩의 사용과 시스템 구현이 비교적 간편하지만 높은 집적도의 구현과 전체 동작의 비동기화가 힘든 디지탈 방식과 집적화가 비교적 쉽고 전체 동작의 비동기화를 자연스럽게 얻을 수 있지만 정밀도와 칩 사용상의 문제가 있는 아날로그 방식이 있으며 또한 두가지 방식의 장점을 융합한 아날로그/디지탈 혼합방식이 있다.

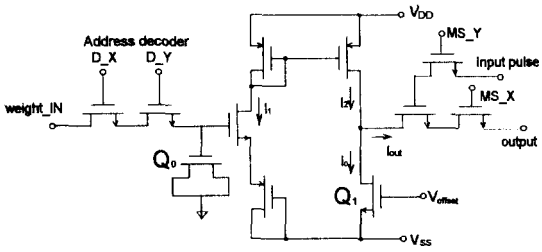
2.2 아날로그-디지탈 혼합 신경망 VLSI - URAN

URAN 칩은 아날로그/디지탈 혼합 구조의 신경망 VLSI이다. 대부분의 동작이 개념적으로는 아날로그 동작을 하며, 특히 디코더를 제외한 나머지 부분은 스위치와 정적 동작 특성을 가지는 MOSFET 저항 연결고리로 이루어졌다. 또한 그림 1의 92 x 92의 기본 모듈 16개로 이루어져, 임의의 외부제어를 통하여 재구성 가능한 특성의 구조이다.



<그림 1> URAN 칩 구조

신경망의 단위 Cell 구조는 그림 2와 같으며 가장 중요한 특성층의 하나인 선형 특성은 0.25%의 해상을 가지고, 256계층 이상이 가능하다.



<그림 2> 단위 Cell 구조

URAN의 연결고리는 양극성을 가지는 전압제어 선형 전류원과 이를 이용한 곱셈기로 동작한다. 즉, 연결고리에 신경망의 연결강도 값을 저장한뒤 펄스를 입력시키면 MOSFET의 triode 영역 선형 채널 저항 특성에 따르는 전류로 전환된다. 따라서 입력에 비례한 펄스를 발생시키면 연결강도와 입력과의 곱셈이 수행된다. 연결강도 값의 전류치 변환은 8 bit 이상의 정밀도를 가짐이 입증되었다. 연결고리의 정밀도 향상이나 구조의 대규모화 구현은 독립적인 양극성 전류원의 wired-OR 확장성으로 실현된다. 특히 기존의 디지탈 방식에서 필수적인 클록이나 동기화 과정이 전혀 필요없다는 점이 확장성이나 변형성에 있어 큰 장점이다.

III. URAN칩을 이용한 음성인식시스템 설계

본절에서는 여러 형태의 URAN칩 중에서 가장 초기에 개발된 256개의 시냅스를 갖는 KTA11을 이용한 음성인식 시스템 설계에 관해 설명한다 [5][6].

이용된 신경망 알고리즘은 3층 구조를 갖는 MLP이며 칩에 인가되는 음성 특징으로는 16개의 필터 뱅크 출력을 가정하였다. 학습은 off line으로 이루어진다. 시스템의 실시간 구현과 연결강도값 적재, 여러가지 제어 동작, 비선형 활성화 작용, 인식결과의 결정 등을 위해서 DSP 보드를 이용한다. 신경망 칩의 모든 인터페이스 회로는 FPGA(Field Programmable Gate Array)의 일종인 Xilinx를 이용해 설계하였다.

3.1 전체 시스템의 구조

KTA11 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.

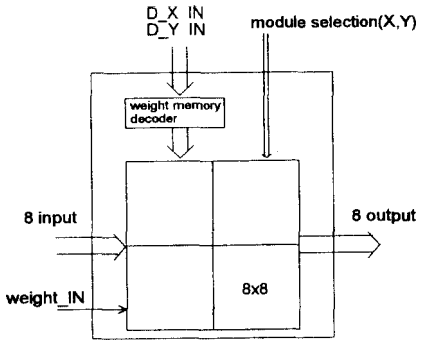
전체 시스템의 개략적인 구성은 그림 4와 같이 PC와 DSP 보드 및 신경망 보드로 구성된다.

3.2 각 시스템의 동작

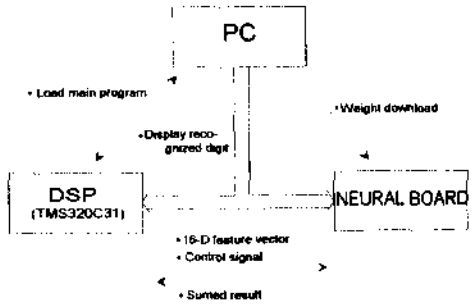
각 시스템의 세부 동작은 아래와 같다.

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.



<그림 3> KTA11의 구조



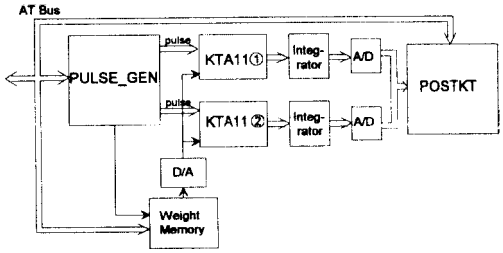
<그림 4> 전체 시스템의 구조

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 제어신호와 함께 신경망 보드로 넘겨진다. 또한 신경망의 각 온닉층에서의 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.

3) 신경망 보드

신경망 보드는 입력 특징벡터를 받아서 매 프레임마다 곱셈과 합을 구하는 과정과 프레임별 결과를 축적하는 역할을 한다. 먼저 PC로부터 해당 프레임의 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 계층의 출력을 구한 뒤 그 값을 DSP 보드에 전송하여 해당 인식단어를 결정하게 된다. 그림 5는 URAN중 가장 간단한 KTA11을 이용한 신경망 음성인식 보드의 구조를 나타낸다. 이때 PULSE_GEN과 POSTKT는 신경망 보드의 제어 신호를 발생시키기 위해서 FPGA로 설계된 2개의 제어칩에 붙여진 이름이다.



<그림 5> URAN을 이용한 신경망 보드의 구조

IV. URAN 칩의 특성을 고려한 음성인식 시스템의 인식 성능 분석

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

4.1 본 연구에 사용된 신경망의 구조

본 연구에 사용된 신경망은 MLP 구조를 가지며 각 층간의 연결고리들은 fully connected 되어 있다. 입력층으로는 17차 벡터의 음성 프레임이 순차적으로 인가되며, 출력층에는 숫자음을 나타내는 node 10개로 이루어져 있다. 각 node들을 연결시켜주는 연결가중치들은 일정한 화자들의 데이타들로부터 학습되어진다. MLP의 학습은 오류 역전파 학습 알고리즘을 이용하였으며 학습률은 0.1, 관성률은 0.9로 하였다. 또한 은닉층 갯수는 30개를 선택하였다.

4.2 URAN 칩을 위한 고려사항

신경망 시스템을 실제로 URAN 칩으로 구성할때의 성능을 검토하기 위해서 URAN 칩에서 제공되는 연결강도의 정밀도, 출력함수등을 이용한 신경망 시스템을 시뮬레이션 한다. URAN 칩 연결강도의 정밀도는 최대 8 bit이므로 우선 8 bit 이하의 정밀도를 가지는 연결강도의 유용성을 검토하기 위해 일반적인 학습과정을 통해 얻어진 부동 소숫점 연결강도의 정밀도를 낮추어 테스트해 보도록 한다. 또한 정밀도를 낮춘 특징벡터와 함께 가장 간단화된 binary 특징벡터를 입력으로 하여 신경망 테스트를 수행한다.

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

4.3 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 x 10 숫자음 x 10명 = 500개) 데이터를 학습하는데 사용하였다. 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 여파된 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 dB가 되도록 원래의 음성 신호와 잡음을 섞었다. 그후 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 1 자리로 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 [7].

4.4 잡음환경에서의 화자독립 숫자음 인식 실험

첫번째 수행한 실험은 sigmoid 비선형 출력함수를 사용한 신경망에 대해 입력과 출력, 연결강도의 정밀도를 각각 소수점 4자리, 2자리, 1자리 등으로 변경하여 학습 및 테스트를 수행하였다. 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 1과 같이 정밀도에 따른 차이는 거의 없었다

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 %

<표 1> Sigmoid 출력함수 신경망의 인식 결과

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.

$$y = a(\sum_i \omega_{ij} O_j) + b, \quad -A \le \sum_j \omega_{ij} O_j \le A \qquad (1)$$

식 (1)로 주어진 선형 출력함수의 정의 영역이 sum - of - weighted - input 값의 범위를 포함할 경우 인식률의 변화는 거의 없었다. 표 2에 a = 0.1, b = 0.5, A = 5인 경우에 대한 화자독립 숫자음 인식률을 비교하였다.

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 2 자리 및

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 %

<표 2> 선형 출력함수 신경망의 인식 결과

1 자리의 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 (1)과 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 (2)와 같은 구간선형 함수를 유도하였다. 소수점 3자리까지 계산했을 때 sigmoid 함수와 식 (2)사이의 MSE는 0이 된다.

$$\begin{aligned} y &= 0.0, && x < -7.6 \\ y &= 0.0087x + 0.066, && -7.7 \le x < -2.2 \\ y &= 0.206x + 0.5, && -2.2 \le x < 2.2 \\ y &= 0.087x + 0.934, && 2.2 \le x < 7.6 \\ y &= 1.0, && 7.6 < x \end{aligned} \qquad (2)$$

식 (2)와 같은 구간선형 출력함수를 사용한 신경망에 대한 학습 시간은 sigmoid 비선형 출력함수를 사용한 경우와 거의 비슷한 (Sun Sparc 10을 사용하여 약 1일) 시간이 걸렸으며, 그 때의 인식결과는 표 3에 나타나 있다.

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 %

<표 3> 구간선형 출력함수 신경망의 인식 결과

표 3에서 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.

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 낮아 질때 신경망의 성능은 저하되지 않았다. 따라서 뉴런칩의 출력함수를 구간 선형함수로 구성한다면 URAN 칩을 이용한 on-chip 학습도 가능할 것이다.

VI. 결론

본 논문에서는 신경망 VLSI와 국내에서 개발된 신경망 VLSI인 URAN에 대해서 알아보았으며, URAN을 이용한 음성인식 시스템을 설계하였다. 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.

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, "신경회로망을 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] 한일송, "신경망 VLSI 기술의 발달과 현재", 한국통신공학회지, 제 9권 제 11호, pp. 47-52, 1992년 11월.

[5] 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.

[6] 이준희, "URAN 신경망 칩을 이용한 숫자음 인식 시스템의 구현에 관한 연구", 한국과학기술원 정보 및 통신공학과 석사논문, 1994.

[7] K. C. Kim and J. W. Cho, "Robust Speech Recognition using Frequency Weighted All-Pole Model Spectrum", *Computer Processing of Chinese & Oriental Language*, vol. 5, no. 3 & 4, pp. 203-216, Nov. 1991.