dSPACE 보드를 이용한 음성인식 명령처리시스템 실시간 구현에 관한 연구

김재웅*, 정원용**
경남대학교

A study on real-time implementation of speech recognition and speech control system using dSPACE board

JaeWoong Kim*, Wonyong Chong**
Kyungnam University
e-mail : ocarina@keri.re.kr, wychong@kyungnam.ac.kr

<요약>

음성은 인간이 가진 가장 편리한 제어전송수단으로 이를 통한 제어는 인간에게 많은 편리함을 제공할 것이다.

본 논문에서는 다층구조 신경망(Multi-Layer Perceptron)을 이용하여 간단한 음성인식 명령처리시스템을 Matlab 상에서 구성해 보았다. 음성인식을 통한 제어의 목적을 위해 화자종속, 고립단어인식기를 목표로 설정하여 연구를 수행하였다. 음성의 시작점과 끝점을 검출하기 위해 단구간 에너지와 영교차율(ZCR)을 이용하였고 인식기의 특징파라미터로는 12차 LPC켑스트럼 계수를 사용하였다. 그리고 신경망의 출력값을 기동, 정지시에 활성화되도록 3개의 계층으로 하였고, 신경망의 뉴런의 개수를 각각 12, 12, 2으로 설정하였다. 먼저 기준음성패턴으로 학습시킨 후에 Matlab 환경하에 동작하는 dSPACE 실시간처리 보드에 변환된 C프로그램을 다운로드하고, 음성을 입력하여 인식 후 dSPACE보드의 D/A컨버터의 출력단에 연결된 DC모터를 기동, 정지제어를 수행하였다. 실시간 음성인식 명령처리 시스템 구현을 통하여 원격제어와 같은 음성명령을 통한 제어가 가능함을 확인할 수 있었다.

Ⅰ. 서론

음성인식 시스템을 구성하는 방법에는 패턴정합에 의한 방법인 DTW(Dynamic time wraping)와 음성의 발생을 통계적으로 모델링한 HMM(Hidden markov model), 그리고 인공신경망을 이용한 방법 등이 있다.

이 중에서 인공신경망은 인간의 뉴런을 단순하게 모델화한 것으로 많은 뉴런의 결합에 의해 적절한 구조를 형성함으로서 패턴분류 문제를 성공적으로 해결할 수 있는 방법이다. 다층구조 신경망은 입력신호들에 대해 교사신호를 적용하여 적절한 목표치를 출력하도록 뉴런들의 결합 가중치를 조절하는 학습을 수행한다.

음성신호에는 많은 신호들이 중복되어 있으므로 효과적인 음성인식을 위해서는 그러한 중복된 정보들을 줄이면서 유효한 정보를 추출하는 특징벡터를 구해야 한다. 그러한 특징벡터로는 현재 데이터와 이전데이터를 적절히 선형 조합하여 다음 시점의 새로운 출력신호를 예측하여 특징벡터 중에서 섞여 있는 두 신호를 분리해 낼 수 있는 켑스트럼이 높은 신호 분해능을 가지므로 LPC켑스트럼을 특징벡터로 사용하여 인식기를 구성하여 보았다.

Ⅱ. 이론

2.1 특징벡터

음성의 시작과 끝점 검출을 위해 단구간 에너지와 단구간 영교차율 함수를 사용하였다. 단구간 에너지 함수는 프레임내의 각 샘플을 자승하

여 모두 더함으로써 근사할 수 있다. 계산을 좀 더 간단하게 하기 위하여 단구간 에너지 함수 대신에 단구간 크기 함수를 쓰면 식(1)과 같다.

$$M_s(m) = \sum_{n=m-N+1=0}^{m} | s(n) | w(m-n) \quad (1)$$

단구간 영 교차율은 프레임 내에서 음성신호가 기준선인 0을 통과하는 횟수를 측정하는 것이다. 무성음의 영 교차율은 무성음 발음 시 난류현상 때문에 유성음의 영 교차율보다 크게 나타나므로 영 교차율은 무성음과 유성음의 구별에 중요한 역할을 한다. 단구간 영 교차율함수는

$$Z_s(m) = \sum_{n=m-N+1=0}^{m} \frac{|sgn\ s(n)| - |sgn\ s(n-1)|}{2} \quad (2)$$

으로 나타내어진다. 여기서

$$sgn\ |s(n)| = \begin{cases} +1, & s(n) \geq 0 \\ -1, & s(n) < 0 \end{cases} \quad (3)$$

이다.

음성의 특징벡터 중에서 선형예측계수(Linear Prediction coefficient)로부터 얻어지는 여러 가지 특징벡터 중에서 켑스트럼이 높은 인식율을 나타낸다는 사실에 근거하여 LPC켑스트럼을 특징벡터로 사용하였다. 선형예측계수 추출은 자기상관 방법인 Levinson-Durbin알고리즘을 이용하였으며 12차 선형예측계수 $\{ a(i),\ 1 \leq i \leq P \}$를 구하였다. 이때 분석모델은 식(4)와 같다.

$$A(z) = 1 - \sum_{i=0}^{p} a(i) z^{-1} \quad (4)$$

LPC켑스트럼 $\{ c(i),\ 1 \leq i \leq P \}$은 선형예측계수로부터 식(5)를 사용하여 계산된다.

$$c(1) = -a(1) \quad (5)$$
$$c(i) = -a(i) - \sum_{k=0}^{i-1} (1 - \frac{k}{i}) a(k) c(i-k), \quad 2 \leq i \leq P$$

여기서 $a(i)$와 $c(i)$는 선형예측계수와 켑스트럼을 나타낸다.

2.2 다층구조 신경망

Matlab상에서 m파일로 코딩된 다층구조 신경망의 구조 중 1층의 내용을 Matlab의 Simulink로 구성한 모델을 그림 1에 나타냈고 신경망의 학습은 D.E. Rumelhart등이 제안한 오류 역전파 알고리즘(error back propagation algorithm)을 사용하였다.

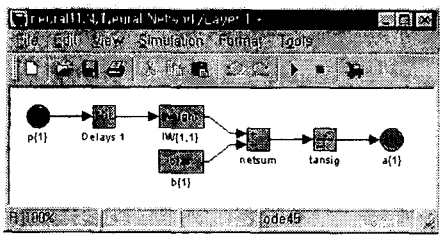

그림1 다층구조 신경망의 1층의 구조

오류 역전파 학습알고리즘은 원하는 출력값과 실제 출력값 사이의 오차를 최소로 하기 위하여 연결강도를 조절하는 Gradient descent방법으로서, 각 노드에서 전달함수를 통하여 실제 출력값을 계산하는 과정과 출력값과 교사값과의 오차를 측정하여 연결강도를 조정하는 과정으로 나눌 수 있다.

출력층과 은닉층에서 각 노드의 오차가 δ_j로 구해지면 각 노드의 연결강도의 변화는 식 (6)의 형태로 구해진다.

$$\Delta W_{ji}(t) = \eta \delta_j o_i + \alpha \Delta W_{ji}(t-1) \quad (6)$$

여기서 학습속도와 직접적인 관계를 가지는 학습율(learning rate) η는 신경망이 수렴하도록 일반적으로 0.01에서 0.25정도의 작은 값으로 주어진다. α는 관성항(momentum)을 나타내며 작은 양수의 값을 주어야 하며 안정성을 강화하면서 학습시간을 개선시키기 위해 연결강도의 변화식에 첨가시켰다.

III. 인식실험

음성은 조용한 실험실환경에서 'Go'와 'Stop'이란 음성을 성인 남성 5인이 20회 발음한 것을

8kH, 8bit로 샘플링하여 수집하였다. 취득된 데이터 중 10회는 신경망의 학습시 사용하였고 나머지 10회분은 신경망의 검사용으로 사용하여 인식율을 검증하였다.

음성의 분석 및 인식의 과정은 아래와 같다. 수집된 음성을 Pre-Empahsis필터를 통과시킨 후, 15 msec의 프레임으로 나누고 10 ms마다 중첩시켜 데이터를 분리한 뒤 각 프레임에 대해 프레임 끝점부분의 오차를 보상하기 위하여 Hamming 창함수를 취한 후. 특징벡터 추출을 위한 프레임의 분석방법으로 자기상관방법의 일종인 Levinson-Durbin 알고리즘을 이용하여 12차 선형예측계수를 구하여 LPC 켑스트럼계수를 구하였다. LPC계수는 신호성분을 잘 표현할 수 있도록 차수를 선정하는 것이 중요하다. 음성을 8kHz로 샘플링했을 때 음성의 공진점이 4개정도 있으므로 선형예측차수 M=10~12정도로 선택함이 좋다. 그림 2에서 보면 24차의 파형은 공진점이 너무 많이 표현되어서 포만트(Formant)주파수를 찾기가 어려움을 알 수 있다.

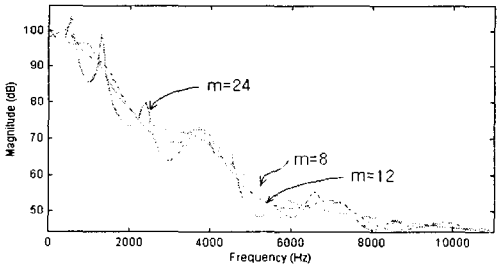

그림2 8차, 12차, 24차 LPC계수의 추출 파형

그림 3은 학습패턴음성입력을 다층구조 신경망에 입력하여 학습을 수행한 결과를 보여주고 있다. 학습율 η는 0.01, 관성항 α는 0.5를 사용하였다. 그리고 각 패턴의 MSE로는 1E-10으로 설정하였다. 1000번 반복을 시켰으나 439회에서 주어진 조건을 만족하여 학습을 멈추었다.

입력층의 노드수는 특징파라미터의 계수에 비례하며 출력층의 노드수는 기동, 정지동작을 수행하기 위해 2개의 뉴런으로 구성하였다.

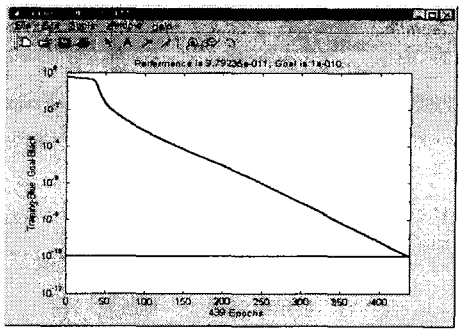

그림3 다층구조 신경망의 학습

실험에 사용된 음성명령인식시스템의 구성도를 그림 4에 나타내었다.

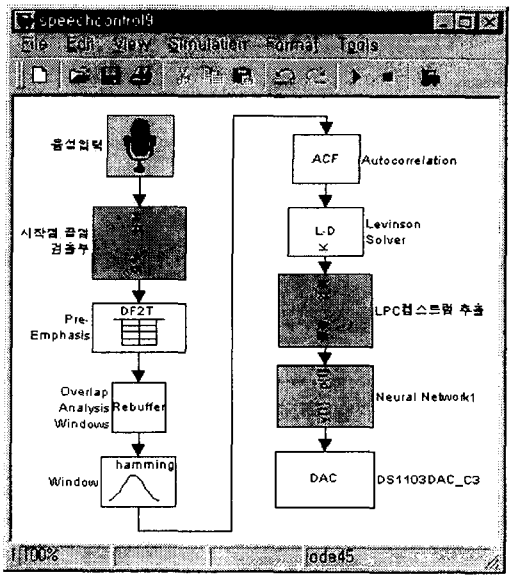

그림4 음성명령인식기의 구조도

dSPACE실시간 처리보드를 통한 실험방법을 잠시 살펴보면 Matlab의 Simulink로 생성된 오프라인 모델을 검증해본 뒤 dSPACE I/O library를 통해 외부신호와 인터페이스하여 Matlab의 RTW에서 생성된 C코드는 dSPACE RTI(Real Time Interface)를 통해 실시간 하드웨어로 다운로딩 된다. 이렇게 다운로드된 프로그램은 PC와는 독립적으로 dSPACE하드웨어에서 실시간으로 계산되어 수행되며 control desk를 이용하여 모니터링하거나 외부장비로 모니터링 하면서 계수들을 조정할 수 있다. 음성명령 인식기의 결과 파형은 그림

을 그림 5에 나타내었다.

'Go'와 'Stop'명령을 번갈아 3번 수행한 결과 파형이다. 이때 'Stop'단어는 인식이 잘 안되었음을 알 수 있다. 신경망의 출력은 0과 1사이의 어떤값도 출력될 수 있으나 0.8이상이면 1로 0.2이하이면 0으로 변환하는 부분을 거쳐 RS-플립플롭을 통과한 출력파형을 나타내었다. 이 출력파형이 1인 부분동안 DC모터가 기동된다.

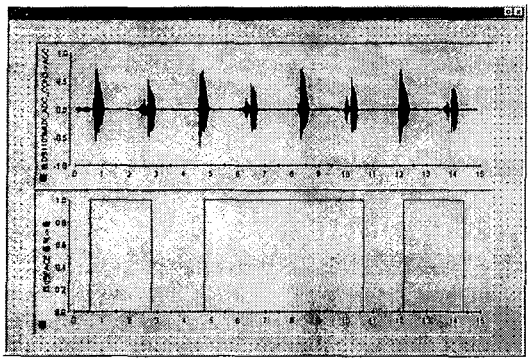

그림5 실시간 음성입력과 결과파형

표 1은 다중구조 신경망을 비실시간과 실시간에서 적용할 때의 인식율을 나타낸 것이다.

인식횟수/입력횟수 *100	Go 음성입력	Stop 음성입력
비실시간	90%	70%
실시간	74%	50%

표1 비실시간과 실시간의 인식율

Ⅳ. 결론

본 연구에서는 다층구조 신경망을 이용하여 고립단어를 대상으로 인식실험을 하였다. 비실시간 인식실험에서는 잡음이 없는 음성신호를 입력하여 인식율이 실시간에 비해 상대적으로 높았으며 실시간 인식은 마이크와 함께 입력되는 잡음이나 음성의 시작점과 끝점 검출의 오류로 인한 오인식이 많았다.

음성명령인식기의 실시간 처리를 위해서는 특징벡터의 크기를 줄여 전체적인 신경망의 계산량을 줄일 필요가 있다. 이에 벡터 양자화(Vector Quantization)등의 방법을 통하여 특징벡터를 줄여 연산량을 감소시키는 것과, 실시간 음성인식시 음성의 시작점과 끝점연산의 부정확함과, 마이크를 통해 음성신호와 동시에 들어오는 잡음성분으로 인하여 인식률이 상당히 저하되었다. 잡음환경에 강한 특징벡터로는 멜켑스트럼(Mel-cepstrum) 계수가 있다고 알려져 있으므로 이의 구현을 통한 잡음환경에 강인한 인식기 구성이 필요하다.

참고 문헌

[1] L. R. Rabiner, R. W. Schafer, "Digital processing of speech signals", pp396~452, Prentice-Hall, 1978

[2] D.P. Mirgan, C.L. Scogield, "Neural Networks and Speech processing", Kluwer Acddemic Publishers, 1991

[3] L. R. Rabiner, Biing-Hwang Juang, "Fundamentals of Speech Recognition", pp 97~121

[4] Sadaoki Furui "Digital Speech Processing, Synthesis, and Recognition", Dekker, 85-125, 1992

[5] D.G. Childers "Speech, Processing and Synthesis Toolboxes", John Wiley & Sons, 2000

[6] Todd K.Moon, Wynn C. Stirling, "Mathematical Methods and Algorithms for Signal Processing", Prentice-Hall, 2000