TDNN과 DTW를 이용한 격리단어 인식

Isolated Word Recognition using TDNN and DTW

황 영 수*

(Young Soo Whang*)

요 약

본 논문에서는 신경 회로망과 DTW를 이용하여 격리 단어 인식을 수행하였다. 인식 대상 단어는 숫자음을 사용하였고, 숫자음에 포함된 음소를 세 부분{ (ㅇ, ㄹ, ㅁ), (ㅅ, ㅊ, ㅍ, ㄱ), (ㅕ, ㅣ, ㅏ, ㅗ, ㅛ, ㅜ) }으로 구분하여 각각의 신경 회로망을 구성한 후, 전체 음소를 인식하기 위하여 세개의 신경 회로망을 합성하였다.

격리 단어 인식은 전단계(합성 신경 회로망) 에서 구한 음소를 이용하여 DTW 기법으로 수행하였다.

ABSTRACT

This paper is a study on the isolated word recognition using neural network and DTW(Dynamic Time Warping). We trained the three group neural networks using the phonetics { (ㅇ, ㄹ, ㅁ), (ㅅ, ㅊ, ㅍ, ㄱ), (ㅕ, ㅣ, ㅏ, ㅗ, ㅛ, ㅜ) } which were included in Korean digits.

In order to recognize total phonemes in Korean digits, we constructed the neural network in a modular fashion by exploiting the hidden structure of previously trained three phonetic subcategory networks.

And isolated words were recognized by the DTW using phonemes found in the previous step.

I. 서 론

디지탈 컴퓨터의 응용 기술과 반도체 기술 및 디지탈 신호 처리 기술이 급격히 발전함에 따라 음성은 인간과 인간 사이의 의사 소통뿐만 아니라, 인간과 기계 사이의 의사 소통을 위한 매개체로서의 역할이 요구되고 있다.

*관동대학교 전자공학과
접수일자 : 1993년 2월 12일

이러한 음성 인식에 대한 연구는 DTW(Dynamic Time Warping) [1], 벡터 양자화(Vector Quantization) [2], HMM(Hidden Markov Model) [3]등을 이용한 연구가 주종을 이루어 왔고, 최근에 인간의 두뇌는 대량의 복잡한 데이터를 병렬 처리할 수 있을 뿐만 아니라 학습 능력이 있다는 사실에 근거하여 새로운 패턴 인식 방법으로 제시된 신경 회로망(neural network)을 이용한 음성 인식에 대한 연구가 활발히 진행되고 있다.[4]-[5].

그러나 이 방법중 시간 변화 정보를 포함하여 격리

단어 인식률이 좋은 DTW 방법은 인식시 계산량과 시간이 많이 소요된다는 문제점을 갖고 있으며, 신경 회로망은 현재 음소 단위를 인식하는데에 많은 연구가 진행되고 있으나 시간에 따른 각 프레임의 정보 변화를 충분히 이용하지 못하기 때문에, 본 연구에서는 격리 단어 인식을 두단계로 구분하여 수행하였다. 즉, 전 단계에서는 신경 회로망을 이용하여 음소 단위로 각 프레임을 인식한 후, 후단계에서는 전단계에서 구한 음소 단위를 사용한 DTW 방법으로 격리 단어 인식을 수행하여, 격리 단어 인식시 계산량과 시간 소요를 적게하면서 인식률을 높이고자 한다.

Ⅱ. 신경 회로망과 DTW

1. 신경 회로망

신경 회로망에서, 생물학적 신경 세포에 해당하는 것이 처리 인자(processing elements)이다. 이 처리 인자들은 수많은 입력 통로(생물학적 수지상 돌기에 비유)를 갖고 있고, 각 입력 통로에 해당하는 값들을 단순한 덧셈 형식으로 수행하게 된다. 여러 결합된 입력 신호들이 전송 함수(transfer function)에 의해 출력에 연결될 것인가를 결정하게 된다. 그림1에 이와같은 신경 회로망의 기본 구성도를 나타내었다.

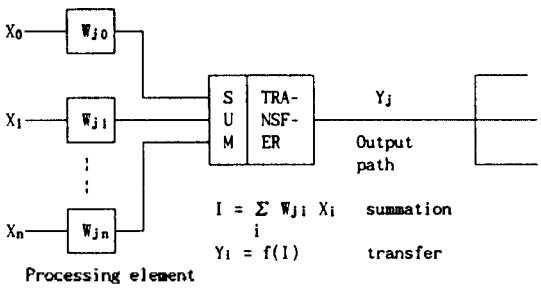

그림 1. 신경 회로망의 기본 구성도

신경 회로망은 위에서 설명한 것과 같이 많은 처리 인자들로 구성되어 있고, 인자들은 그림2에 나타낸 것과 같이 여러 층(layer)으로 분리되어, 각 층의 인자들은 불규칙적으로 서로 연결되어 있다. 그림2에서 입력 버퍼(input buffer)는 회로망에 데이터를 입력시키는 작업, 출력 버퍼(output buffer)는 주어진 입력에 대한 회로망의 응답을 유지시키는 작업을 하게 되며, 은닉층(hidden layer)은 입력 버퍼와 출력 버퍼를 서로 구분시켜 주는 역할을 하게 된다.

이와 같은 회로망의 작업은 우선 각 가중치(weight) 들을, 각 입력에 대해 원하는 출력을 얻기 위하여, 학습 작업중에 적응시키야 한다.

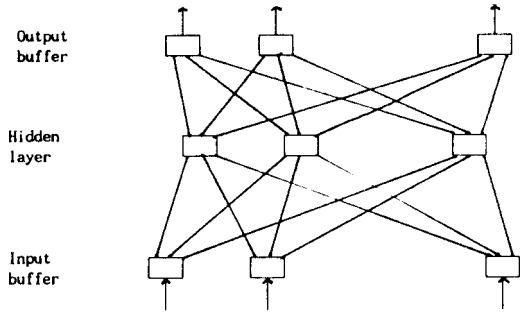

그림 2. 신경 회로망 처리 인자들의 구조

2. DTW

DTW 알고리듬은 시험 패턴과 기준 패턴을 시간축상에 최적이 되도록 배열한 후, 이 최적 변형 경로를 통한 최적 거리를 얻어내는 방법이다.

DTW 알고리듬은 2차원 $d_i(m, n)$ 평면상에서 식을 한 제한 조건을 만족하는 d(1, 1)에서 d(M, N)까지의 최적 경로를 구하는 방법이다. 임의의 점(m(j), n(j))까지 축적된 거리를 다음과 같이 정의할 수 있다.

$$C_i(m, n)=\sum_{j=1}^{J} d_i(m(j), n(j)) * w(j)$$

여기에서 (m(j), n(j)), j=1, 2, ……, J는 주어진 경로이고, w(j)는 각 경로에 따른 가중치이다. 그러면 최적 경로는 d(M, N)에서의 축적된 거리 $C_i(M, N)$을 최소화 하는 경로로

$$D_i(T, R_i)=\min_{\text{path}} C_i(M, N)$$

로 표시된다. 최적 경로를 구하는 과정에서 제한 조건들은 다음과 같은 목적으로 주어진다. 즉, 부분적으로는 경로 기울기의 범위를 제한하며, 전체적으로는 경로의 허용 영역을 제한한다.

Ⅲ. 본 연구에서 수행한 격리 단어 인식 시스템

본 연구에서는 전단계에서 신경 회로망을 이용하여 음소를 인식한 후, 후단계에서 DTW 방법으로 격리 단어 인식을 수행하였다. 그림3에 본 연구에서 수행한 격리 단어 인식 시스템을 나타내었다.

그림3의 전단계에서 사용한 신경 회로망은, 신경 회로망 입력에 시간 지연을 갖는 그림4와 같은 신경 회로망을 구성하여 각 프레임의 음소 인식을 수행하였다.

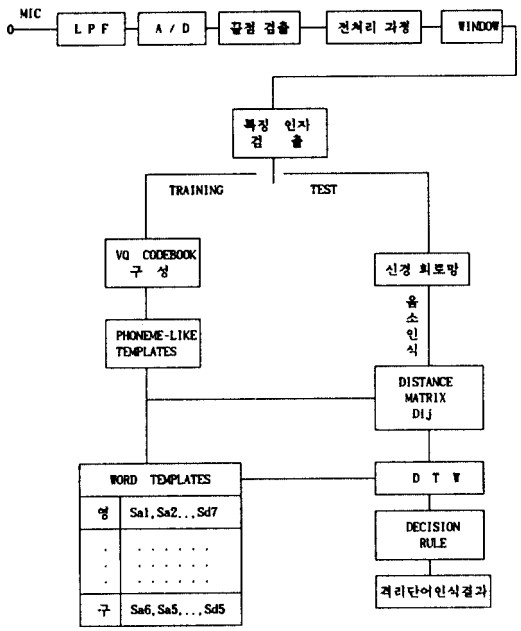

그림 3. 본 연구에서 수행한 음성 인식 시스템 블록도

그림4와 같이 입력단에 시간 지연을 갖는 인자를 입력시킴으로서, 한 순간에 해당하는 특징 인자뿐만 아니라 전 프레임과 현 프레임에 해당하는 인자들의 비선형 관계를 이용하여 음소를 인식할 수 있다. 또한 여러 음소를 한 신경 회로망에 학습을 시킬 경우, 신경 회로망에 연결된 가중치의 수가 증가됨에 따라 학습시 많은 시간을 소요하기 때문에, 그림4에 나타낸 것과 같이 인식할 음소들중 유사 특성을 갖는 음소들끼리 분할시켜 적은 음소들로 구성된 신경 회로망을 우선 학습시킨후, 여러 신경 회로망을 한개의 신경 회로망으로 합성시켜 학습시키는 방법을 이용하였다.

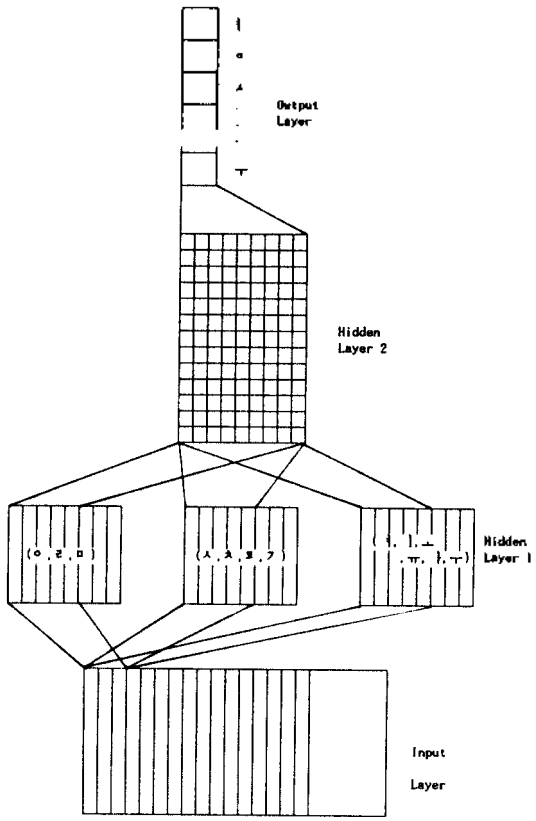

그림 4. 유사 음소별로 구성한 후 합성한 신경 회로망

그림3의 후단계에서 DTW를 수행할때 인식 계산량을 줄이기 위하여, 그림3의 음소 형태의 코드워드 상호간의 거리 행렬과 이 음소 형태 코드워드를 이용한 단어 표준 패턴을 구성하였다. 여기에서 음소 형태 코드워드는 대표 단어들을 벡터 양자화 하여 구성한다. 단어 표준 패턴은 대표 단어에 음소 형태 코드워드를 이용하여, 각 시간열의 특징 벡터와 거리를 구한 후, 가장 작은 거리를 갖는 음소 형태 코드워드를 그 단어내의 프레임을 나타내는 대표값으로 선정하였다. 이와 같은 방법을 이용할 경우, 일반 DTW 방법에 비해 기억 용량과 계산 시간을 줄일 수 있다.

Ⅳ. 실험결과 및 고찰

1. 데이터 베이스의 구성

데이터 베이스를 구성하는 첫 단계로서 20대 남성 화자를 선택한 후, 숫자음 영(0) 부터 구(9)까지를

15번 만복 발음하여 릴 테이프(reel tape)에 녹음하였다.

이와 같이 릴 테이프에 녹음한 후, 3.7KHz의 대역폭을 갖는 저대역 통과 필터(low pass filter)로 고주파 성분을 제거하였다.

저대역 통과 필터로는 B & K사의 주파수 분석기(Type 2200)를 사용하였다.

필터를 통과한 음성 데이터는 IBM PC에 부착된 16bit A/D 변환기를 사용하여 디지탈화 하였다. 이때 샘플링 주파수는 10KHz로 하였으며, 녹음된 음성이 A/D 변환기의 입력 범위에 맞도록 증폭기를 이용하여 이득(gain)을 조정하였다.

이와 같은 과정을 그림5에 나타내었다.

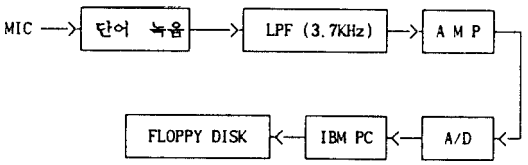

그림 5. 데이터 베이스를 구성하는 과정

2. 인식결과

신경 회로망을 이용하여 음소를 인식하기 위하여 사용한 학습 데이터는 화자 1인이 각 숫자음을 발음한 데이터중에서, 숫자음에 포함된 음소 부분(ㅕ, ㅇ, ㅣ, ㄹ, ㅅ, ㅁ, ㅗ, ㅟ, ㄱ, ㅊ, ㅍ, ㅜ)을 15 프레임씩 구분하여 14차 켑스트럼 계수를 구하였으며, 인식시에는 학습시 포함되지 않은 숫자음 90개를 이용하였다. 이때 각 음소의 시간 지인 특성을 표현시키기 위하여, 현 프레임과 2개의 전 프레임을 학습과 인식시 이용하였다.

표1에 전체 음소 13개에 대한 신경 회로망을 구성하여 인식한 실험 결과를 나타냈으며, 인식 결과는 84.8%의 인식률을 얻었다.

또한 유사 특성 음소 즉, 무성음(ㅅ, ㅊ, ㅍ, ㄱ), 유성 자음(ㅇ, ㄹ, ㅁ)과 모음(ㅕ, ㅣ, ㅏ, ㅗ, ㅟ, ㅜ)에 대해 각각의 신경 회로망을 구성하여 인식한 실험 결과를 표2에 나타냈으며, 각각 90.2%, 93.1%, 87.4%의 인식률을 얻었다. 이 실험결과 신경 회로망을 이용하여 음소를 인식할 경우, 인식할 음소의 수가 많을수록 인식률이 저하된다는 것을 알 수 있었다.

표3에 표2에서 이용한 3개의 신경 회로망을 합성하여 전체 13개의 음소를 인식한 결과를 나타냈으며, 인식 결과는 82.4%의 인식률을 얻었다. 그러므로 이와 같이 음소들을 분할하여 신경 회로망을 구성한 후 전체 음소를 인식할 경우, 분할된 신경 회로망을 합성하면, 전체 음소에 대한 신경 회로망을 사용하는 것보다 적은 학습 시간을 소요하면서 비슷한 인식 결과를 얻을 수 있었다.

격리 단어 인식시 DTW의 음소 형태 표준 패턴 음소는 전단계 신경 회로망 학습시 이용한 데이터중에서 음소당 각각 10개를 선정하였으며, 격리 단어 인식시 사용한 데이터는 신경 회로망 학습시 이용하지 않은 데이터중에서 각 격리 숫자음당 10개의 데이터를 사용하였다.

표4에 격리된 단어 인식 결과를 나타내었으며, 인식결과는 99%의 인식률을 얻었다.

[표 1] 전체 음소에 대한 인식 결과 (1 개의 신경회로망 이용)

	ㅕ	ㅇ	ㅣ	ㄹ	ㅅ	ㅏ	ㅁ	ㅗ	ㅟ	ㄱ	ㅊ	ㅍ	ㅜ
ㅕ	52							7	10				
ㅇ		62											
ㅣ			67										
ㄹ		7		65			10						
ㅅ					57						11	8	
ㅏ						72							
ㅁ		3		8			62						
ㅗ								46					7
ㅟ	16							12	50				
ㄱ										72			
ㅊ					8						61		
ㅍ					7							64	
ㅜ	4		5					7	12				65

[표 2] 유사 특성 음소로 구분된 신경 회로망을 이용한 음소 인식 결과

(a) 모 음

	ㅕ	ㅣ	ㅏ	ㅗ	ㅟ	ㅜ
ㅕ	62		2		4	1
ㅣ		62		1		
ㅏ		4	67			2
ㅗ			3	61		5
ㅟ	7			4	64	
ㅜ	3	8		6	4	64

(b) 무 성 음

	ㅅ	ㄱ	ㅊ	ㅍ
ㅅ	66	1	4	4
ㄱ		70		2
ㅊ	4		65	
ㅍ	2	1	3	66

(c) 유성 자음

	ㅇ	ㄹ	ㅁ
ㅇ	67		4
ㄹ	2	68	2
ㅁ	3	4	66

이와 같은 결과는 DTW를 사용함으로서 인식 후, 각 프레임의 시간에 따른 정보 변화량을 고려한 결과라고 생각한다.

그러나 인식 대상 숫자음이 신경 회로망 학습시 사용하지 않은 다른 화자의 격리 숫자음을 인식할 경우에는, 인식률이 급격히 저하되는 것을 보였다.

[표 3] 전체 음소에 대한 인식 결과 (3개의 신경회로망 합성한 신경회로망 이용)

	ㅎ	ㅇ	ㅣ	ㄹ	ㅅ	ㅏ	ㅂ	ㅗ	ㅜ	ㄱ	ㅅ	ㅍ	ㅜ
ㅎ	50							6	11				
ㅇ		60		1									
ㅣ			61			2			1				
ㄹ		9		66			8						
ㅅ					60						9	6	
ㅏ			3			65							4
ㅂ		3		5			64						
ㅗ			1			5		43					8
ㅜ	17							14	47				
ㄱ										69			
ㅅ					6					1	63		
ㅍ					6					2		66	
ㅜ	5		7					9	13				60

[표 4] 격리 단어 인식 결과

	영	일	이	삼	사	오	육	칠	팔	구
영	10									
일		10	1							
이			9							
삼				10						
사					10					
오						10				
육							10			
칠								10		
팔									10	
구										10

V. 결 론

본 논문에서는 신경 회로망과 DTW를 이용한 격리 단어 인식에 관한 연구를 수행하였다. 음소를 인식하기 위하여, 시간 지연 특성을 갖는 데이터를 신경회로망의 입력 인자로 사용하여, 정적 상태가 아닌 동적 상태의 프레임들 비선형 변화를 이용하였다. 또한 신경 회로망을 이용하여 패턴을 인식할 경우, 인식할 패턴의 수가 증가함에 따라 인식률이 저하되고, 학습 시간이 많이 소요되기 때문에, 인식 대상의 수를 적게한 신경 회로망들을 합성한 신경 회로망을 이용하여 인식할 음소수를 증가시켰다. 이와같은 신경 회로망으로 인식된 음소를 이용하여 격리 단어 인식을 수행하기 위하여, DTW 기법을 사용하여 각 프레임간의 시간 정보 변화량을 고려하였다.

실험 결과 유사 특성을 갖는 음소들끼리 분할시킨 후, 이 분할된 신경 회로망들을 합성한 신경 회로망을 이용한 음소 인식률은 82.4%, 이 음소 결과를 DTW에 결합시켜 구한 격리 단어 인식률은 99%의 결과를 얻어, 학습 시간과 인식 시간의 소요를 적게하면서 높은 격리 단어 인식 결과를 얻을 수 있었다.

앞으로 학습 시간 소요 문제와 학습시 여러 화자의 데이터를 사용하여, 화자 독립 음소 인식을 좀 더 연구한 후, 이 결과를 토대로 신경 회로망을 이용한 음소 인식 방법을 격리 단어 인식에 적용시킬 예정이다.

참 고 문 헌

1. G. M. White and R. B. Neely, "Speech Recognition Experiment with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoust. Speech and signal Processing, Vol. ASSP-24, pp. 183-188, April, 1976.
2. Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design, "IEEE Trans. on Com, Vol. COM-28, Jan., pp. 84-95, 1980.
3. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, Vol. 77, Feb 1989.
4. A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang," Phoneme Recognition using Time-Delay Neural Network," IEEE Trans. Vol. ASSP-37, Aug., 1989.
5. R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Mag, Vol. 4, pp. 4-22, 1987.
6. D. J. Burr, "Experiments on Neural Net Recognition of Spoken and Writen Text," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-36, pp. 1162-1168, July 1988.
7. 백준호, 김유신, 손경식, "은닉층 뉴우런 추가에

의한 역전파 학습 알고리듬," 전자공학회논문집 제29권 4호, pp. 58-65, 1992.
8. A. Waibel and K. F. Lee, Reading's in Speech Recognition, Morgan Kaufmann Pub. 1990.

▲황 영 수

1982년 2월 : 연세대학교 전자공학과 졸업(공학사)
1984년 2월 : 연세대학교 대학원 전자공학과 졸업 (공학석사)
1990년 2월 : 연세대학교 대학원 전자공학과 졸업 (공학박사)
1989년 9월 ~ 현재 : 안동대학교 전자공학과 조교수

본 논문은 1991년 한국학술진흥재단의 학술조성비에 의해 연구되었음.