論文93-30B-12-3

신경망과 구문분석을 이용한 한국어 연결 숫자음 인식

(Connected Korean Digit Recognition Using Neural Networks and Lexical Analysis)

李 鍾 碩*, 李 商 郁*

(Jong Seok Lee and Sang Uk Lee)

要 約

본 논문에서는 신경망과 숫자음의 구문분석을 이용한 한국어 연결숫자음 인식 시스템을 제안하였다. 제안된 방법은 먼저 단독 발음된 숫자음절에서 위치정보에 의해 추출된 기준음소 데이타를 이용하여 음소분류 신경망의 학습을 수행하며, 학습된 신경망을 이용하여 매 프레임을 음소로 레이블링 한다. 그리고 레이블링된 프레임들을 결합하여 음소 세그먼트를 형성하고, 이들을 숫자음절 결합규칙에 의해 음절로 결합한다. 이와같이 결합된 음절은 숫자음절 결정조건에 의해 인식을 수행하며, 결정조건을 만족시키지 못한 음성 세그먼트는 다음 단계에서 음절 결정 신경망에 의해 인식하게 된다.

본 논문에서는 5명이 10번씩 발음한 단독숫자음을 이용하여 신경망 학습을 수행하였으며, 제안된 인식 시스템의 성능을 평가하기 위하여 이들 5명이 발음한 30종류 숫자열을 이용하여 컴퓨터 시뮬레이션을 수행하였다. 실험 결과 95.6%의 숫자 인식률, 82%의 숫자열 인식률을 얻었다.

Abstract

In this paper, we propose a connected Korean digit recognition system employing neural networks and lexical constraints of the Korean digits. In the proposed recognition system, firstly, each frame of digit string is labelled by phoneme classification neural networks, which are trained with the reference phoneme segments extracted from an isolated digit based on the position information. And, the frame labels are combined with each other for constructing the phoneme segments. Then, these segments are combined to form a digit candidate using the digit combination rules. The digit candidate is decided based on the condition for digit decision. If the condition is not satisfied, the digit candidate is further recognized using the digit decision neural network in the next step.

In our approach, the neural networks are trained with 10 isolated digits uttered by 5 male speakers. To investigate the performance of the proposed recognition system, an intensive computer simulation on the 30 connected digit strings uttered by 5 male speakers is performed. The simulation result indicates that 95.6% digit recognition rate and 82% digit string recognition rate are provided by the proposed Korean digit recognition system.

*正會員, 서울 大學校 공과대학 제어계측 공학과
(Dept. of Control and Instrumentation Eng., Seoul Nat'l Univ.)
接受日字 : 1992年 7月 14日

I. 서 론

인간의 음성을 컴퓨터를 이용하여 인식할 수 있다면 인간과 기계와의 획기적인 정보 전달 수단으로 이용될 수 있을 것이다. 이러한 음성인식의 가장 기본

적인 방법은 고립적으로 발음된 단어를 인식하는 것으로 DTW(dynamic time warping)[1]나 HMM (hidden Markov model)[2] 등에 의해 좋은 결과를 얻고 있다. 그러나 인간은 많은 음성정보를 전달하기 위해서 단어들을 연결시켜 발음을 한다. 단어들이 연결될 경우 단어와 단어 사이에 묵음이 존재하는 고립단어 인식과는 달리 몇가지 어려운 문제점이 대두된다. 첫째, 단어와 단어가 연결되면서 음운현상 및 조음현상이 발생되어 음의 변형이 일어나게 된다. 둘째, 단어가 연결되면서 단어간의 경계가 불명확해져 경계를 찾기가 어렵게 된다. 셋째, 발음습관 및 음운현상에 의한 단어 길이의 변화가 심하다. 넷째, 인식 대상 단어수가 늘어날 경우 계산량은 기하급수적으로 증가한다. 이러한 연결단어를 인식하는 대표적인 방법으로 패턴 비교를 이용하는 two-level DP법[3], level building DP법[4], one stage DP법[5] 등과 통계적 특성을 이용하는 HMM[6] 등이 사용되고 있다.

그러나 인간이 음성을 이해하기 위해서 사전에 습득한 대량의 지식과 통계적, 구조적 해석에 의한 고도의 인식기능을 사용하듯이, 연결단어 인식에서 인식 성능을 높이기 위해서는 인식 대상 단어의 음성학적, 음향학적 표현에 대한 정보와 이들이 문맥상에서 어떻게 결합되는가의 정보를 이용해야 한다. 음성의 특징을 표현하는 방법은 구문분석을 통한 음성학적 기호나 특징에 의해 표현하는 방법과 확률적인 분포나 기준 패턴에 의해 표현하는 방법이 있다. 전자의 구문분석적 표현 방법은 지식 데이타베이스를 갖는 시스템의 행동에 대한 추론을 가능하게 한다. 즉, 예를 들어 음소를 레이블링하여 인식하는 시스템의 경우 음성의 특징이 외적으로 표현되기 때문에 언어학적 표기에 대응시켜 레이블링된 음소열 정보의 이해 및 정정을 가능하게 한다. 또한 음성학적, 음향학적 제약조건을 사용함으로써 계산량의 감소를 얻을 수 있다. 그러나 인식 단어수가 증가할 경우 지식 데이타베이스를 구축하기가 어려워진다. 이러한 표현방법은 단어간의 결합뿐만 아니라 음소의 결합시에도 적용이 가능하여 발음망(pronunciation network)을 이용하는 방법[7][8], fuzzy inference rule을 이용하는 방법[9] 등이 제안되었다. 후자의 표현 방법은 자동적인 clustering과 파라메타 예측기법을 사용할 수 있다. 즉, 음성학적, 음향학적 개념에 음성 데이타를 직접 대응시키지 않고 통계적 특성을 사용하여 인식할 수 있으나 통계적 특성을 얻기 위해서는 많은 학습데이타가 필요하다. 이러한 표현 방법에는 DTW[1], HMM[2] 등이 있다. 현재 후자의 표현 방법이 전자의 표현 방법 보다 좋은 결과를 얻고 있으나 전자의 경우 음소단위의 정확한 표현과 이들의 결합에 대한 지식 데이타베이스를 잘 구축할 경우 훌륭한 결과를 기대할 수 있다.

한편 60년대 초반 한때 중단되었던 신경망이 80년대 후반 부활되면서 우수한 패턴 분류 능력 때문에 음성인식 분야에도 많은 시도가 이루어지고 있고, 좋은 결과도 얻고 있다.[10] 신경망은 간단한 처리요소들이 병렬적으로 구성되어 데이타를 분산, 병렬 처리함으로써 실시간이 가능한 계산속도를 얻을 수 있고, 학습에 의한 적응력이 있어 많은 데이타를 학습시킬 경우 훌륭한 패턴분류 능력을 발휘할 수 있다.[14]

본 논문에서는 이러한 신경망의 패턴분류 능력과 연결 숫자음을 구성하는 구문분석을 이용한 복수화자의 연결 숫자음 인식에 관하여 고찰해 보고자 한다. 인식 대상 어휘는 음성 다이얼링 시스템, 음성 데이타 입력 시스템 등에 이용이 가능하도록 숫자음으로 제한하였다. 신경망의 우수한 패턴분류 능력을 이용하여 음의 기본 단위인 음소를 표현하고 이들 음소들이 음절로 결합되는 규칙을 설정하여 숫자음절을 인식하는 시스템을 제안하였다. 이렇게 시스템을 구성할 경우 앞에서 언급한 연결단어 인식시 발생하는 문제점들이 해결될 수 있으리라 기대된다. 즉, 숫자음절들이 연결되어 발음될 경우 음운현상 및 조음현상이 발생하여 음의 변형이 일어나게 되는데 신경망은 분류하고자 하는 패턴을 학습된 패턴들 사이에서 상대적인 유사도에 의해 판별해 내는 능력이 있기 때문에 약간 왜곡된 패턴들도 훌륭히 분류해 낼 수 있다. 또한 이러한 음의 변형이 단어간 또는 음소간의 경계 분리를 어렵게 하지만 제안된 방법에서는 매 프레임별로 음소를 레이블링 하여 해결하고 있다. 같은 음소들이 연속적으로 레이블링 되다가 왜곡이 심한 경계 부분에서 부분적으로 레이블링이 안되거나 다른 음소로 레이블링이 되지만, 인접한 프레임들의 레이블링값과 음소 결합규칙을 이용하면 해결이 가능해진다. 발음습관 및 음운현상에 의한 단어 길이의 변화는 분류된 음소 세그먼트를 직접적으로 기준 패턴과 비교하지 않고 분류된 음소 레이블을 사용하기 때문에 음소의 길이에 대한 정보를 사용함으로써 해결할 수 있다. 계산량 측면에서는 신경망을 학습시키기 위해서는 많은 양의 데이타 패턴과 계산을 요하나, 인식시에는 신경망에 의한 패턴분류도 간단한 계산으로 가능하고 음절 결합규칙도 논리로 구성되어 있기 때문에 패턴비교에 의한 인식 방법과는 비교가 되지 않을 정도로 적다.

인식을 위한 기준 패턴 작성 방법으로는 고립단어

에서 추출하는 방법[3] [4] [5], 조음현상이 포함되도록 여러 경우에 대해 숫자를 연결시켜 발음한 후 경계를 분리하여 기준 패턴을 작성하는 방법[11] [12]이 사용되고 있다. 본 논문에서는 신경망을 사용하여 패턴을 분류하기 때문에 연결단어 발음시 발생하는 음의 변형은 신경망의 상대적인 유사도 판별에 의해 가능하므로 고립적으로 발음된 숫자음절에서 음소를 분리하여 기준 패턴으로 사용하였다. 음소를 정확히 분리하기 위해서는 수작업으로 경계를 일일이 구분하여 사용해야 하지만 앞에서도 언급하였듯이 신경망의 우수한 패턴분류 능력 때문에 숫자음절에서 간단한 위치정보를 사용하여 음소를 분리하여 신경망을 학습시켰다. 이렇게 학습된 신경망을 사용하여 음소들을 분류하고, 분류된 음소들을 음절 결합규칙에 의해 숫자음절로 결합한 뒤 두단계로 숫자음을 인식할 수 있는 인식 시스템을 제안하였다. 제안된 시스템은 5 명의 8 자리 전화번호 음성 데이타를 사용하여 성능을 평가하였다.

Ⅱ. 음성인식 시스템

1. 구성

연결 숫자음 인식 시스템은 그림 1과 같이 구성된다. 음성신호가 입력되면 전처리및 특징 파라메타를 추출한 후 에너지를 이용하여 음성구간과 묵음구간을 검출한다. 각각의 음소로 학습된 13개의 음소분류 신경망은 매 프레임이 어느 음소와 가장 유사한가를 각각의 신경망의 출력을 통해 분류해 낸다. 이 출력들을 세그먼트 결합부에서 인접한 프레임들의 결합논리에 의해 음소 세그먼트로 결합한 후, 음절 결합부에서 음절 결합논리에 의해 숫자음절로 결합하여 1차 결정부로 보낸다. 1차 결정부에서는 음성/묵음 정보와 결합된 숫자음절 정보를 이용하여 결정논리에 의해 숫자음절을 결정한다. 만일 1차 결정부에서 숫자음이 결정되지 못하면, 결정되지 못한 음절부분을 분리하여 시간축 상의 비틀림(time warping) 보정과정을 거치고 음절결정 신경망을 이용하여 숫자음을 다시 결정하게 된다. 이와 같이 본 논문에서 제안하는 연결 숫자음 인식 시스템은 크게 전처리 및 특징 추출부, 음소분류 신경망 및 음절결정 신경망, 음절결합부 및 결정부의 세 부분으로 나눌 수 있다.

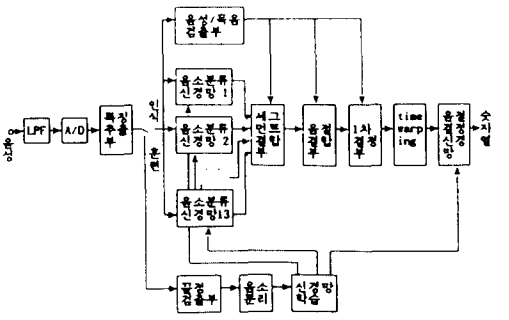

그림 1. 연결 숫자음 인식 시스템 구성도

Fig. 1. The block diagram of the connected digit recognition system.

2. 전처리 및 특징 추출부

음성신호를 차단 주파수 4.5 kHz인 저역필터를 통과시킨 후, 10 kHz로 표본화하고 16 비트로 양자화하여 16 msec 분석구간에서 특징 파라메타를 추출한다. 특징 파라미타는 묵음구간의 검출 및 유성음/무성음의 특징을 포함할 수 있도록 에너지, 영교차율 및 인접한 샘플간의 상관 정도를 나타내는 1차 반사계수(first order reflection coefficients)와, 음소간의 특징을 표현하는 파라메타로서 선형예측 분석방법을 통하여 얻은 LPC 계수로 부터 얻어지는 10 차 LPC 켑스트럼 계수를 사용한다.[13] 따라서 특징 파라메타는 한 프레임당 13개의 원소를 갖는 벡터로 구성된다. 이들 특징 파라메타들은 신경망에 사용하기 위해서 각기 설정된 최대값으로 나누어 줌으로써 -1과 1 사이값으로 정규화시킨다.

3. 음소분류 신경망 및 음절결정 신경망

1) 음소분류 신경망

인식하고자 하는 연결 숫자음성은 매 프레임 마다 특징 파라메타를 추출하여 그 프레임이 어떤 음소 특성을 갖는가를 음소분류 신경망을 사용하여 분류하게 된다. 본 논문에서 선정한 인식대상어는 전화번호에 사용되는 연결 숫자음으로 "공", "일", · · · , "구"로 구성되는 7 개 숫자음에 연결음 "에"를 포함시켜 8자리로 구성하였고, 이들은 각각 표 1과 같이 13 개의 음소로 분류를 할 수 있다. 이중 "육"의 경우는 더욱 세분화 될 수 있으나 짧게 발음되고, 특성이 급격히 변하고, 뒤에 묵음이 존재하는 등 그 자체로 인식이 가능하여 별도로 분류 한다. 음소분류 신경망은

표 1. 음소 구성표

Table 1. Phoneme table of the Korean digit.

초성	ㄱ ㅅ ㅊ ㅍ
중성	ㅗ ㅣ ㅏ ㅜ ㅔ
종성	ㅇ ㄹ ㅁ
기타	육

한 신경망이 한 음소를 분류할 수 있도록 구성하였기 때문에 같은 형태를 갖는 13개의 신경망으로 이루어진다. 신경망의 구조는 다층인식자(multi-layer perceptron) [14]를 사용하고 은닉층은 2층, 입력은 현재 프레임 및 앞 뒤의 한 프레임씩 세 프레임으로 구성하여 39개의 입력노드를 갖도록 하고 출력은 현재의 프레임이 그 음소의 부류에 속하는지 아닌지를 나타내도록 한개의 노드로 한다.

2) 음절결정 신경망

그림 1에서 음소분류 신경망의 출력들을 결합하여 음소 세그먼트를 만들고 이 음소들의 구성에 의해 1차 결정부에서 숫자음절을 결정하게 되는데, 만일 결정하지 못할 경우에는 다음 단계로 결정이 미루어진다. 이 두번째 단계의 결정에 사용되는 신경망이 음절결정 신경망이다. 연속된 숫자열들은 대부분 1차 결정부에서 결정이 되지만 주어진 조건을 만족하지 못하여 결정을 못할 경우 음성이 존재하는 영역을 분리하여 시간축 상의 일정한 길이로 정규화 시킨 후 음절결정 신경망에 입력시켜 판단을 한다. 이때 음성 세그먼트의 앞 부분에서 초성자음이 검출되면 초성자음 부분에 비중을 두어 그림 2와 같이 정규화 시킨

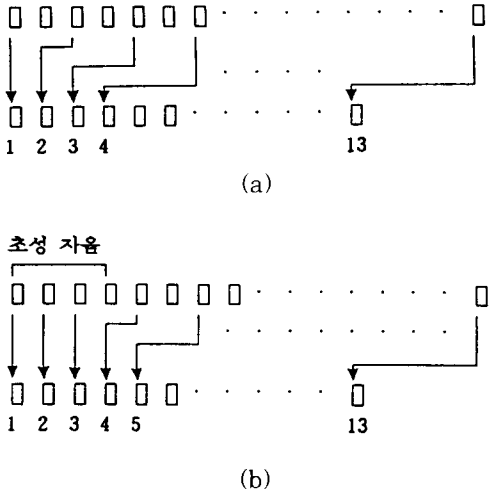

그림 2. 음절결정 신경망의 학습용 데이타의 비선형 정규화
(a) 자음으로 시작하지 않는 경우
(b) 자음으로 시작하는 경우

Fig. 2. Non-linear time warping of the training data for digit decision neural network.
(a) The case which does not start consonant.
(b)The case which starts consonant.

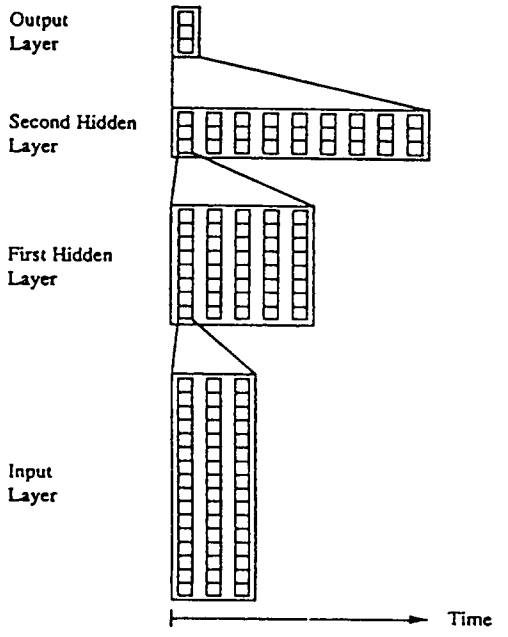

그림 3. TDNN의 구조

Fig. 3. The architecture of the TDNN.

다. 보통 전화번호는 발음습관상 국번호 3자리와 연결음 "에"뒤에 상당기간 묵음을 두고 그뒤에 번호 4자리를 발음한다. 따라서 음성구간중에서 장시간의 묵음을 검출한 후 이 묵음의 앞뒤로 4자리의 숫자음을 검출하게 된다. 만일 음성 세그먼트가 일정 길이보다 긴 경우 두 음절인가 확인하여야 한다. 결정된 음절과 결정못한 음성 세그먼트의 갯수가 장시간의 묵음의 앞뒤로 4자리씩 8자리 인가를 확인하고, 부족할 경우 종성 음소 세그먼트와 초성 음소 세그먼트가 만나는 부분을 찾아 두 부분으로 나누어 주고 만일 만나는 부분이 없으면 중간에서 나누어 준다. 음절결정 신경망은 음성 세그먼트가 시간축 상의 정규화 과정에서 왜곡이 발생하기 때문에 시간축상의 변이를 포함할 수 있는 그림 3의 구조를 갖는 TDNN(time delay neural network) [10]을 사용한다. 신경망의 입력은 각 숫자음을 모두 13 프레임으로 정규화시켜서 형성하고, 두개의 은닉층과 10개의 숫자음 및 "에"를 나타내도록 11개의 출력 노드를 갖도록 구성한다.

(3) 신경망의 학습

- 음소분류 신경망의 학습

음소분류 신경망의 학습용 데이타는 단독 발음된 숫자음에서 음소별로 채취한다. 각 음소의 경계를 정확히 검출할 수 있으면 바람직 하겠으나 일일이 수작

표 2. 음소분류 신경망의 학습용 음소 데이타의 채취 위치 및 갯수

Table 2. The position and number of the training data for phoneme classification neural networks.

음 소		위치	갯수
초	ㄱ	공, 구의 앞 부분	6
	ㅅ	삼, 사 〃	6
성	ㅊ	칠 〃	3
	ㅍ	팔 〃	3
중	ㅏ	사의 중간 및 끝 부분	3
	ㅣ	이의 전 부분	3
	ㅗ	오의 전 부분	3
성	ㅜ	구의 중간 및 끝 부분	3
	ㅔ	에의 전 부분	3
종	ㄹ	일, 칠, 팔의 끝부분	9
	ㅁ	삼의 끝 부분	3
성	ㅇ	공의 끝 부분	3
	육	육의 중간 및 끝 부분	3

업으로 처리하지 않는 이상 어렵다. 그러나 실제 인식시에 매 프레임마다 음소분류를 수행하므로 대부분 같은 음소들이 연속되고 경계 부분에서 몇 프레임의 음소분류 에러가 발생할 수 있다. 이 에러는 분류된 프레임 레이블의 평활화 및 음절결합 규칙에 의해 극복이 가능하므로 음소의 대략적인 위치 정보를 이용하여 표 2와 같이 학습용 데이타를 추출한다. 한 음소 영역에서 학습용 데이타는 세 쌍씩 취하게 된다. 즉, 표 2에서 위치가 앞 부분이면 음성 데이타의 앞에서부터 세 프레임의 데이타를 한 프레임씩 이동시키면서 세 쌍을 취하고, 중간 및 끝 부분 또는 전 부분은 주어진 범위에서 양 끝 및 중간에서 세 쌍을 취한다. 이렇게 할 경우 한 사람이 11음절을 한번씩 발음한 데이타에서 음소 데이타를 추출하면 표 2와 같이 51개의 학습용 패턴이 채취된다. 본 연구에서는 5명이 10번씩 발음한 데이타를 이용하므로 51 × 5 × 10 = 2550개의 학습용 데이타가 생성된다. 이 데이타로 13개의 음소 분류 신경망을 학습시키게 되는데 해당 음소 신경망에 대응되는 음소 데이타는 원하는 출력을 "1" 로 나머지 학습용 데이타는 "0"으로 주어 학습 시킨다. 즉, "ㄱ" 음소를 분류해내는 신경망의 학습용 데이타는 "ㄱ"에 해당하는 6 × 5 × 10 = 300개의 데이타는 원하는 출력을 "1"로, 나머지 2250개의 데이타는 "0"으로 하여 순서를 임의로 섞어 학습 시킨다.

- 음절결정 신경망의 학습

음절결정 신경망은 11개의 음절 데이타로 TDNN을 학습시키는 과정이다. 길이가 서로 다른 11개의 음절 데이타를 시간축 상에서 13프레임(0.208 sec)으로 모두 정규화시켜 학습 시킨다. 그런데 자음으로 시작하는 음절들은 자음이 음절결정에 중요한 역할을 하므로 그림 2와 같이 초성자음 부분에 비중을 두어 비선형 정규화 시킨다. 초성자음이 없는 음성 세그먼트는 전 프레임에 걸쳐 균일하게 13프레임으로 정규화 시키고, 자음으로 시작하는 음성 세그먼트는 자음 부분이 최소한 세 프레임이 되도록 정규화 시킨다.

Ⅲ. 음절결합부 및 음절결정부

1. 음소 세그먼트 결합부

음소 세그먼트 결합부의 입력은 13개의 음소분류 신경망의 출력열들이다. 이들 출력열들은 해당 음소가 존재하는 영역에서 그 음소 신경망의 출력이 "1"의 값을 갖는다(출력값이 문턱값 이상이면 "1"로 한다). 그러나 해당 음소의 해당 영역에서 "1"이 되지 않는 프레임이 존재할 수도 있고, 다른 음소도 동시에 "1"이 될 수도 있다. 이들은 평활화(smoothing)를 하고 같은 음소들 끼리 결합하여 음소 세그먼트로 해 주어야 한다. 평활화는 자음의 경우 한 프레임만 "1"로 존재할 경우 제거시키고 연속되다가 한 프레임이 끊어진 경우 연결시킨다. 모음의 경우는 두 프레임 기준으로 수행한다. 비슷한 영역에서 두개 이상의 음소 세그먼트가 생성 되는 경우도 있는데 이 경우 어떤 세그먼트가 옳은지 결정할 필요가 없고 모두 음소 세그먼트로 취급한다. 왜냐하면 음절결합 과정에서 옳은 세그먼트만 음절로 결합되기 때문이다.

2. 음절결합부

음소 세그먼트들이 결합되면 이들을 조합하여 숫자 음절을 결합하여야 한다. 이들의 결합은 표 3의 결합 규칙에 의해 수행한다. 표 3에 중성이 생략되면서 결합되는 규칙이 있는데 (예: 공 => ㄱ+ ㅇ) 이는 신경망 훈련 데이타의 작성시 정확한 음소분리가 이루어지지 않고 중성 부분이 조금씩 초성과 종성으로 섞여 들어간 패턴이 인식된 경우이다. "육"의 경우 앞에 짧은 "이"가 올 수 있고 뒤에 묵음이 존재한다. "에"의 경우는 앞의 세자리 뒤에 붙는데 전화번호 발음 습관상 뒤에 긴 묵음이 따른다. 또한 "이"와 "오"의 경우 일정 길이 이상이 되면 두번 발음된 것으로 보고 중간 부근 에너지가 가장 작은 부분에서 나누어 준다. 숫자가 연결 발음 되면서 발생되는 대표적인

음운현상은 연음, 경음화, 유성음화 등이 있다. 이 중 경음화(/s/ -> /s'/), 유성음화(/k/ -> /g/)의 경우는 음의 특성이 크게 변하는 것이 아니므로 학습된 데이타중 상대적으로 가장 유사한 패턴으로 분류해 내는 신경망의 패턴분류 능력상 별도의 처리없이 인식이 가능하고, 연음의 경우는 음소의 길이는 변하지만 음소의 조합은 같으므로 인식이 가능하다. 또한 "육육"이 "융육"으로 발음되는 경우는 표 3에 "융"의 처리규칙을 둠으로써 인식이 가능해 진다.

표 3. 숫자음절 결합규칙

Table 3. The combination rules of the Korean digits.

공	ㄱ + ㅗ(ㅜ) + ㅇ ㄱ + ㅇ
일	이 + ㄹ ㄹ
이	이
삼	ㅅ + ㅏ + ㅁ ㅅ + ㅁ
사	ㅅ + ㅏ
오	오
육	이(짧은) + 육 + 묵음 육 + 묵음
칠	ㅊ + ㅣ + ㄹ ㅊ + ㄹ
팔	ㅍ + ㅏ + ㄹ ㅍ + ㄹ
구	ㄱ + ㅜ(ㅗ)
에	에 + 묵음(발음 습관상)
융	ㅣ(짧은) + 육 + ㅇ 육 + ㅇ

3. 1차 결정부

1차 결정부는 음절결합부에서 음소 세그먼트들의 결합에 의해 형성된 음절영역을 음절결정 모델을 이용하여 숫자음으로 결정하여 주는 부분이다. 음소 세그먼트들이 표 3과 같은 규칙에 의해 숫자음절로 결합되면 그 숫자음을 음절결정 모델에 의해 결정한다. 숫자음의 결정은 결합형태에 따라 표 4와 같이 자음이 있는 그룹과 없는 그룹, 기타의 세 그룹으로 나누어 다음과 같이 수행한다. 그림 4에서 먼저 초성자음

표 4. 인식 그룹 분류표

Table 4. The table of the digit recognition group.

그룹 Ⅰ	공, 구, 삼, 사, 칠, 팔
그룹 Ⅱ	이, 오, 일
그룹 Ⅲ	육, 에

여부를 판단한 후 초성자음이 있으면 그룹 I의 형태 판단을 하여 결정하고, 초성자음이 발견되지 않으면 그룹 III, 그룹 II의 순으로 결정한다. 그룹 III을 그룹 II보다 먼저 결정하는 이유는 특징이 보다 명확하기 때문이다. 각각 그룹내의 음절 결정은 음절결정 모델과의 유사도 판단에 의해 결정한다.

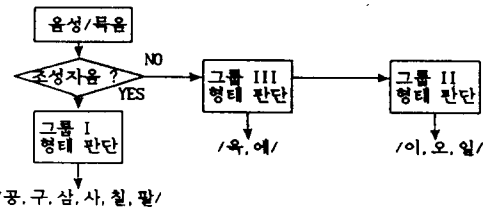

그림 4. 1차 결정부의 결정 순서

Fig. 4. The flow chart of the primary decision procedure.

음절결정 모델은 인식 대상 숫자음절을 구성하는 음소 세그먼트들에 의해 결합된 형태에 관한 정보를 가지고 있도록 구성한다. 즉, 그림 5와 같이 각 숫자음절에 대해 모델을 설정할 수 있다. 이 모델은 학습된 음소분류 신경망과 이들의 학습에 사용한 숫자음을 이용하여 만든다. 학습에 이용한 10번씩의 숫자음 데이타를 13개의 음소분류 신경망에 입력시켜 각 음소분류 신경망의 출력값들을 10번씩 누적시키면 그림 5와 같은 음절결정 모델을 생성할 수 있다. 모델내에서 음절을 구성하는 음소 세그먼트 영역은 그림 5와 같이 큰값을 갖고 나머지 영역은 0에 가까운 값을 갖게 된다. 이 음절결정 모델의 형태를 보면 음소분류 신경망이 얼마나 정확히 학습되었는가를 알 수 있고, 유사한 음소의 경우에는 작은 값이지만 출력이 존재함을 알 수 있다. 그림 5의 음절결정 모델은 각각 "공", "일", · · · , "구", "에" 의 11개 모델을 나타내고, 한 모델에서는 시간축을 20프레임으로 고정시켰고, 음소축은 차례대로 "ㄱ", "ㅅ", "ㅊ", "ㅍ", "ㅇ", "ㄹ", "ㅣ", "ㅁ", "ㅏ", "ㅗ", "육", "ㅜ", "ㅔ" 의 13개 신경망의 출력값에 대응된다. 이 음절

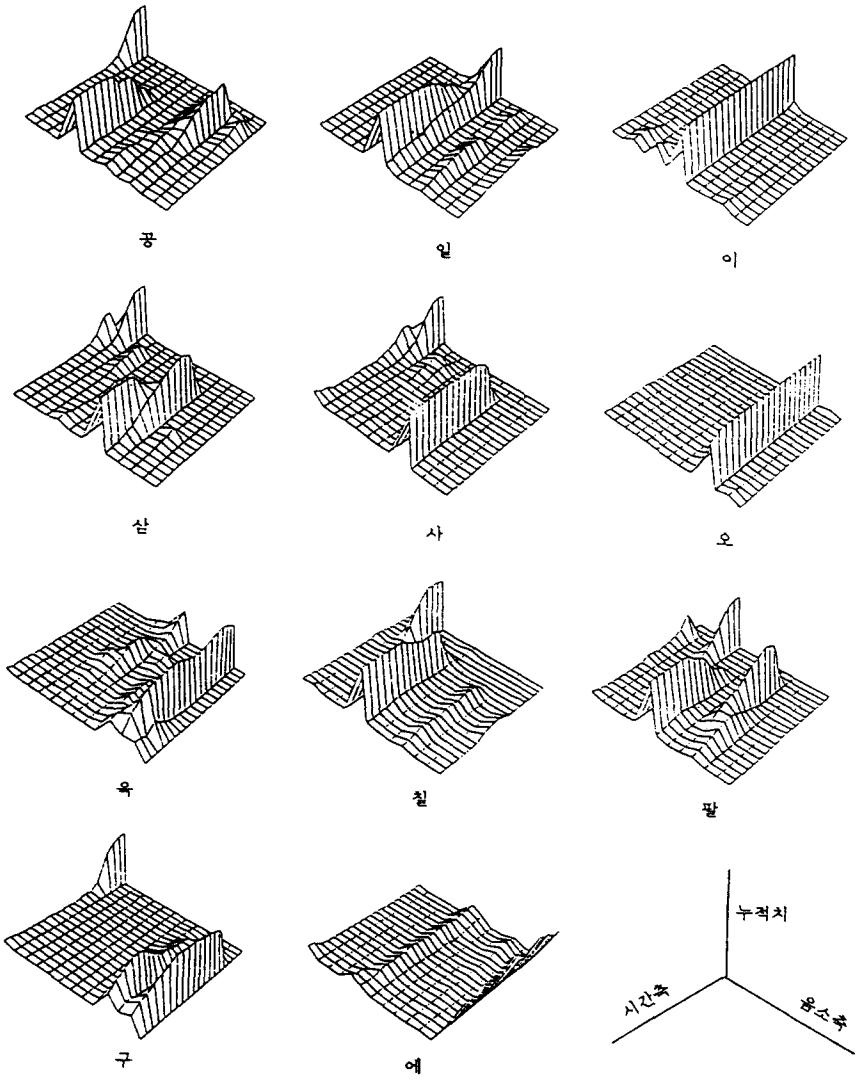

그림 5. 음절 결정 모델
Fig. 5. syllable decision model.

결정 모델을 이용하여 숫자음절로 결합된 영역, 즉 숫자음 후보에 대해 인식을 수행한다. 숫자음 후보는 음절결정 모델과 같은 길이가 되도록 선형 비틀림(linear warping) 과정을 거쳐 20프레임으로 만들고, 11개의 음절 결정 모델과의 일치도를 각각 측정한다. 매 프레임 마다 음절결정 모델의 각 음소의 누적치 X_{ij}와 이에 대응하는 숫자음 후보의 음소분류 신경망의 출력 Y_{ij}를 곱하여 13개의 음소 중 최대값을 찾은 후 20개 프레임의 값을 다음 식과 같이 합산 하면 숫자음 후보와 음절 결정 모델과의 일치도를 계산할 수 있다.

$$\sum_{i=0}^{19}[MAX_{1\le j\le 13}\ X_{ij} \bullet Y_{ij}]$$

이때 일치도가 일정 문턱값 (최대 일치도 × 0.25) 보다 크고, 11개 모델 중 가장 높은 음절결정 모델의 숫자를 인식 결과로 한다. 대부분의 경우 이 과정에서 숫자음으로 결정되지만 음이 아주 짧은 경우, 자음 부분이 아주 약하여 자음 신경망으로조차 검출이 안되는 경우 등은 결정하지 않은 상태로 Ⅱ.3.2절의 음절결정 신경망으로 보낸다.

Ⅳ. 실험 및 결과

인식실험에 사용된 데이타는 성인 남자 5명의 화자가 8자리 전화번호를 30종류 발음한 150개의 숫자열

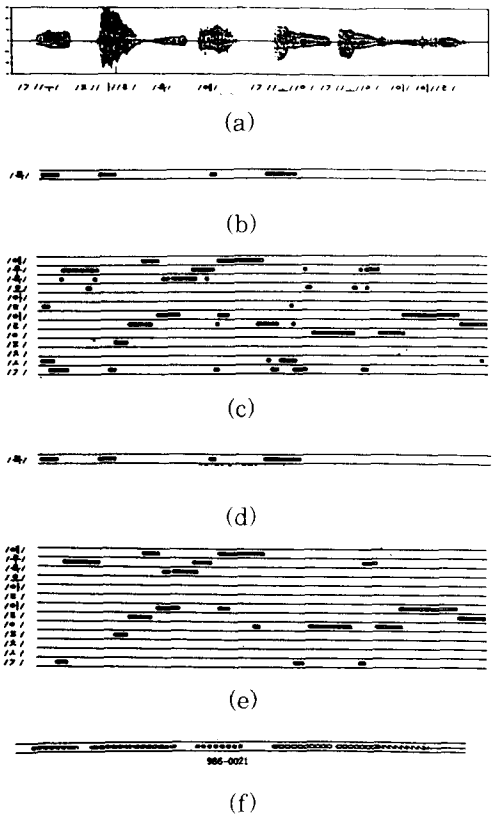

그림 6. 숫자열 결정 과정의 예
(a) 시간축 상에서의 음성파형
(b) 묵음 검출 결과
(c) 음소분류 신경망의 출력
(d) 묵음 검출의 평활화 결과
(e) 음소분류 신경망 출력의 평활화 결과
(f) 인식 숫자열 (986-0021)

Fig. 6. An example of the digit string decision procedure.
(a) The waveform of the speech.
(b) The results of the silence detection.
(c) The outputs of the phoneme classification neural networks.
(d) The smoothed results of the silence detection.
(e) The smoothed outputs of the phoneme classification neural networks.
(f) Recognized digit string(986-0021).

로 1200개의 숫자음절로 이루어져 있다. 데이타는 주변잡음이 없는 조용한 실험실에서 채취하였다. 신경망의 훈련에 사용된 데이타는 한 사람이 11음절(공, 일, · · · , 구, 에)을 10번씩 발음한 110개의 음절 데이타 5명분을 사용하였다. 실험에 사용한 음소 분류 신경망은 13개 모두 같은 형태의 다층 인식자로 입력노드는 39개, 첫번째 은닉층의 노드는 30개, 두번째 은닉층의 노드는 10개, 출력 노드는 1개로 구성하였다. 음절 결정 신경망은 TDNN으로 입력층은 한 프레임당 13개의 노드로 구성하고 두 프레임의 시간지연을 두었고, 첫번째 은닉층은 20개의 노드에 5 프레임의 시간지연을, 두번째 은닉층은 11개의 노드에 5 프레임의 시간지연을 두었고, 출력노드는 11개로 구성 하였다. 각 신경망의 학습시에는 momentum rate 및 learning rate는 각각 0.3 으로 하였고 총 에러값이 0.001 이하가 될때 까지 학습하였다. 그림 6에 한 숫자열의 결정과정의 예를 도시하였다. (a)는 "구팔육에 공공이일"의 시간축상의 파형을 나타낸다. 이 신호에서 프레임별로 특징 파라메타를 추출한 후 에너지를 이용하여 묵음을 검출한 결과가 (b)에 도시되어 있다. 출력값이 "1"인 경우만 "O"의 심볼로 표현하였다. 또한 이 특징 파라메타들이 13개의 음소분류 신경망에 입력되어 판단된 결과가 (c)에 도시되어 있다. 이 결과들을 세그멘트 결합부에서 평활화하여 (d)와 (e)의 결과를 얻었다. (e)에서 음소 분류 신경망의 출력이 "1"이더라도 묵음 구간으로 분류된 부분은 심볼로 표현하지 않았다. (e)의 결과는 음절결합 규칙에 의해 결합되고 숫자음절이 결정된 후 최종적으로 "986-0021"의 숫자열을 내주게 된다. (e)의 결과를 보면 음절결합시 중성이 없이 연결되는 경우가 많이 발생되는데 4.2절에서도 설명한 바와 같이 신경망 훈련시 엄밀한 음소 분류가 이루어지지 않았기 때문이고, 이 문제는 표 3의 음절결합 규칙에 의해 해결되고 있다. 또 (e)의 "에" 음절영역의 앞 부분과 뒷부분에 "이" 와 "ㄹ"의 음소 세그먼트가 존재하지만 음절결정 조건을 만족시키지 못해 결정되지 못하고 있다. 또한 "육"의 경우 "육"신경망 출력의 앞부분에 "이"가 중복적으로 존재하지만 결합규칙에 의해 육으로 결합되는 것을 알 수 있다. 전체적인 인식실험 결과가 표 5 및 표 6에 제시되어 있다. 1차 결정부에서 숫자음을 잘못 인식한 경우를 포함하여 결정된 숫자음은 1200 음절중 1139 음절로 94.9%의 결정률을 보였다. 즉 2차 결정부인 음절결정 신경망으로 보내지는 음절은 5.1% 이다. 이중 올바르게 결정된 인식률은 표 5와 같이 93.8% 이다. 결정하지 못한 부분을 분리하여 음절 결정 신경망으로 결정하였을 때의 인식률은 95.6% 이다. 이때 8자리로 구성되는 숫자열의 인식률은 82% 이다. 숫자열 인식률이 매우 낮은 것은 8 숫자중 한자만 틀려도 오인식으로 간주 하

였기 때문이다.

표 5. 숫자 인식률
Table 5. Recognition rates of the digit.

화자	1	2	3	4	5	평균인식률(%)
1차 결정부	94.2	90.0	95.0	95.4	94.2	93.8
음절결정 신경망	96.3	92.5	96.3	96.7	96.3	95.6

표 6. 숫자열 인식률
Table 6. Recognition rates of the digit string.

화자	1	2	3	4	5	평균인식률(%)
인식률	83.3	86.7	80.0	80.0	80.0	82.0

본 논문의 실험결과는 기존 알고리듬에 의한 실험결과나 여타 논문의 결과와 비교 검토가 이루어져야 하나, 음성인식은 실험 대상 음성 데이타에 따라 인식성능에 큰 차이가 있어 직접적인 비교가 불가능하다. 영어의 경우에는 표준 데이타가 있어 서로 인식성능을 비교하고 있으나 국내에는 표준 데이타가 현재 존재하지 않고 있다. 국내에서 이에 대한 연구와 대책이 무엇보다도 필요하다고 하겠다. 따라서 본 논문에서 사용한 데이타를 이용하여 기존의 알고리듬을 실험하여 결과를 비교 검토해야 할 필요성이 있어, 향후 성능 개선시 병행할 계획이다.

V. 결론

단독 발음된 숫자음절에서 위치정보에 의해 대략적으로 추출된 음소성분 데이타로 학습된 음소분류 신경망과 숫자음을 구성하는 구문분석을 이용하여 5 명의 연결 발음된 숫자음 데이타를 인식하는 한국어 연결 숫자음 인식 시스템을 구성하였다. 연결 숫자음 데이타는 음소로 구분되고, 이 음소들을 음절결합 규칙에 의해 음절로 결합한 후 숫자음절 결정 조건에 의해 인식하고, 인식하지 못한 음성 세그먼트를 분리하여 음절결정 신경망에 의해 숫자음절을 인식해 내도록 하였다. 실험 결과 숫자음 같이 어휘가 한정되어 음소의 갯수가 적고 음운현상이 명확할 경우, 단독 발음된 숫자음에서 채취된 음소 데이타로 학습된 신경망은 훌륭한 음소분류 및 음소영역 구분 능력을 발휘할 수 있으며, 연결된 음성의 발성시 발생하는 음운현상도 신경망의 패턴분류 능력과 간단한 음절결합 규칙으로 극복될 수 있음을 알았다. 또한 기준 패턴도 단독 발음된 숫자음에서 간단히 추출하였기 때문에 인식 시스템 구현시 유리하리라 사료된다. 본 논문에서는 5명의 화자에 대해서 실험하였으나 앞으로 화자수가 늘어날 경우 효율적인 신경망의 재학습 방법, 문턱값들의 적응화 방안이 강구되어야 할 것이다.

參考文獻

[1] H. Sakoe, S. Chiba, "Dynamic programming algorithm for spoken word recognition," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-26, pp.43-49, Feb. 1978.

[2] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden Markov models to speaker independent, isolatd word recognition," *AT&T Bell Lab. Technical Journal*, vol. 62, no. 4, pp. 1075-1105, Apr. 1983.

[3] H. Sakoe, "Two-level DP-matching - A Dynamic programming based pattern matching algorithm for connected word recognition," *IEEE Trans. Acoustis., Seech, Signal Processing*, vol. ASSP-27, pp. 588-595, Dec. 1979.

[4] C. S. Myers and L. R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans., Acoustic., Speech, Sgnal Processing*, vol. ASSP-29, pp. 284-297, Apr. 1981.

[5] H. Ney, "The use of a one-stage dynamic programming for connected word Recognition," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-32,pp. 263-271, Apr. 1984.

[6] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High performance connected digit recognition, using hidden Markov models," *IEEE Proc. ICASSP-88*, pp. 119-122, New York, Apr. 1988.

[7] D. Klatt, "Scriber and LAFS : Two new approaches to speech analysis," in *Trends in Speech Recognition*, W. Lea, Ed. Englewood Cliffs, NJ : PrenticeHall,

1980.

[8] G. Kopec and M. Bush, "Network-based isolated digit recognition using vector quantization," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASP-33, pp. 850-867, Aug. 1985.

[9] P. Demichelis, R. De Mori, P. Laface, and M. O'Kane, "Computer recognition of plosive sounds using contextual information," *IEEE Trans. Acoustic., Seech, Signal Processing*, vol. ASSP-31, pp. 359-377, Apr. 1983.

[10] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time delay neural networks," *IEEE Trans. Acoustic., Speeh, Signal Processing*, vol. ASSP-37, pp. 328-339, March 1989.

[11] L. R. Rabiner, J. G. Wilpon, Ann M. Quinn, and Sandra G. Terrace, "On the application of embedded digit recognition," *IEEE Trans. Acoustic., Speech, Signal Processing*, vol. ASSP-32, pp. 272-279, Apr. 1984.

[12] 김민성, 안승권, 차신, 이종석, "한국어 연속 숫자음 인식," 음성 통신 및 신호 처리 workshop, 1990.

[13] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, New York: Spriger-Verlag, 1976.

[14] Y. H. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addision Weley, 1989.

著者紹介

李鍾碩(正會員)

1960年 7月 14日生. 1983年 2月 서울대학교 제어계측공학과 졸업 (학사). 1985年 2月 서울대학교 대학원 제어계측공학과 졸업 (석사). 1985年 3月 ~ 현재 금성사 중앙연구소 재직. 1991年 2月 ~ 현재 서울대학교 제어계측공학과 박사과정

李商郁(正會員) 第 22券 第 1號 參照

현재 서울대학교 제어계측공학과 교수