신경 회로망을 이용한 연속 음성에서의 keyword spotting 인식 방식에 관한 연구

A study on the Method of the Keyword Spotting Recognition in the Continuous speech using Neural Network

양 진 우*, 김 순 협**
(Jin-Woo Yang*, Soon-Hyob Kim**)

요 약

본 논문은 keyword spotting 기술을 이용한 247개의 DDD 지역명을 인식 대상으로 하여 화자 독립의 한국어 연속 음성 인식을 위한 시스템을 제안하였다. 적용된 인식 알고리즘은 음성에서 시간축의 변화와 스펙트럼의 왜곡을 흡수할 수 있는 모델로 DP와 MLP로 구성된 동적 프로그래밍 신경회로망(DPNN)을 사용하였다. 이와 같은 실험을 위해 단어 모델을 만들고 이에 대한 단어 모델을 keyword 모델과 non-keyword 모델로 구분하여 성능을 향상시킬 수 있도록 하였다. 또한 잘못된 결과를 출력시키지 않기 위해서 후처리 과정을 두고 실험을 하였다.

실험 결과, 단독어에 대한 화자 종속 실험은 93.45%의 결과를 보였고, 단독어에 대한 화자 독립 실험은 84.05%의 실험 결과를 보였으며, 가장 중요한 간단한 대화체 문장의 keyword spotting 실험은 화자 종속으로 77.34%의 결과를 보였으며, 화자 독립 실험은 70.63%의 결과를 얻었다.

ABSTRACT

This research proposes a system for speaker independent Korean continuous speech recognition with 247 DDD area names using keyword spotting technique. The applied recognition algorithm is the Dynamic Programming Neural Network(DPNN) based on the integration of DP and multi-layer perceptron as model that solves time axis distortion and spectral pattern variation in the speech. To improve performance, we classify word model into keyword model and non-keyword model. We make an experiment on postprocessing procedure for the evaluation of system performance.

Experiment results are as follows. The recognition rate of the isolated word is 93.45% in speaker dependent case. The recognition rate of the isolated word is 84.05% in speaker independent case. The recognition rate of simple dialogic sentence in keyword spotting experiment is 77.34% as speaker dependent, and 70.63% as speaker independent.

Ⅰ. 서 론

본 연구는 핵심주제어 추출 기술을 이용하여 247개의 DDD 지역명을 인식 대상으로 한 화자독립의 한국어 음성 인식에 관한 연구이다. Keyword spotting이란 음성인식의 한 분야로서 컴퓨터가 사람의 음성을 입력받아 이 음성에 미리 정해진 특정 단어(keyword) 또는 복수 개의 단어들 중 어느 것이 포함되어 있는지의 여부를 찾아내고 이 단어를 식별해 내는 작업을 의미한다. keyword spotting은 자연스러운 연속음성으로부터 꼭 필요한 정보(keyword)를 추출해 냄으로써 비록 최소한의 수준이기는 하지만 사람과 컴퓨터간의 자연스러운 의사소통을 가능케 해 준다는 점에서 사용자의 편리함이 강조되는 추세와 더불어 그 역할의 중요성이 점차 증대되고 있다[6], [11]. 따라서, 본 논문은 단어를 기본단위로 하는 DPNN (Dynamic Programming Neural Network)을 이용하여 keyword spotting 시스템을 구현하였다. 음성인식을 하는데 가장 어려운 것은 시간축의 왜곡과 스펙트럼의 왜곡이다. 기존의 시간축의 왜곡을 해결하는 수학적인 방법은 DP 정합 방법이다. 반면에 스펙트럼의 변화는 처리하기가 힘들다. 신경회로망은 패턴 인식을 하는데 우수한 결과를 보여준다. 따라서 음성에서 스펙트럼의 변화를 흡수하는데 신경회로망을 이용한다. 이와같이 각각의 장

*춘천 기능 대학 전자 기술학과
Electronic Technology Department, ChunChon Polytechnic Collage
**광운대학교 컴퓨터공학과, 신기술 연구소
Dept. of Computer Engineering & Institute of New Technology, KwangWoon University

접수일자 : 1996년 5월 6일

심을 가지고 있는 DPNN 모델을 음성인식 시스템에 사용하였다.

Ⅱ. 제안된 keyword spotting 시스템

2.1 단어 인식을 위한 신경회로망(DPNN)

동적 프로그래밍 신경회로망, DPNN(Dynamic Programming Neural Network)은 MLP와 DP를 이용한 음성인식 신경회로망이다[1]. 음성 인식을 하는데 가장 어려운 것은 화자의 신체적인 상태와 시간에 따라 음성 신호의 크기가 심하게 변하고 시간 지연(Time-delay) 현상이 빈번함에 따라 단어의 파형이 크게 변한다는 것이다. 따라서 음성 인식 시스템을 구현하기 위해서는 음성 신호로부터 특징되는 패턴을 안정적으로 추출할 수 있는 능력이 있어야 한다.

또한 패턴인식 능력에 있어서 음성 신호로부터 포만트 분포와 같은 정적인 특징과 포만트 천이와 같은 동적인 특징을 추출할 수가 있어야 한다. 이런 Invariance 특성과 패턴인식 능력을 이용할 수 있다면 화자독립 연속 음성인식을 실현 할 수 있을 것이다. 말하자면 화자독립 특성은 Speaker Invariant 기능으로부터 가능해지는데 이것은 신경망의 일반화 특성 때문이다. 그리고 연속 음성인식은 Time Invariant 특성 때문에 가능해진다.

이러한 기능을 고루 갖춘 신경회로망이 본 논문에서 제안한 동적 프로그래밍 신경회로망(DPNN)이다. 다시 말해서, DP 정합 방법으로는 시간축의 왜곡을 처리하였고, 신경회로망은 스펙트럼의 변화를 흡수하는데 사용되었다. 이와 같은 기능을 갖는 제안된 신경회로망의 구조는 그림 2.1에 잘 나타나 있다.

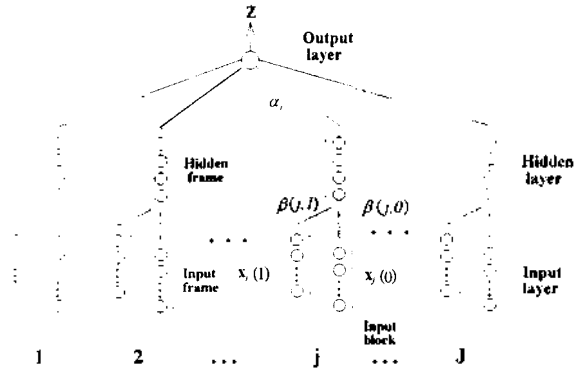

그림 2.1 동적 프로그래밍 신경회로망

위의 그림 2.1에서 신경회로망은 입력 패턴이 주어진 범주(category)에 속하는지 아닌지를 결정하는 분류기의 역할을 한다. 즉, 네트웍의 출력 또는 출력 유닛으로부터의 신호가 입력 패턴과 네트웍을 대표하는 범주 사이의 유사성을 나타내도록 한다는 것이다. 그리고 시간축 j를 갖는 프레임 구조의 입력층과 은닉층으로 구성되어진다. 각 입력 프레임은 K개의 입력 유닛(j, k)를 갖고, 각 은닉 프레임은 L개의 은닉 유닛(j, l)을 포함하고 각 은닉 프레임(j)로 τ만큼의 시간이 지연된 입력 프레임이 연결된 입력 블럭(j)로 나타낼 수 있다.

은닉 유닛(j, l)의 출력은 다음과 같다.

$$y(j, l) = f(net(j, l)) \tag{2.1}$$

여기서, $f(\cdot)$은 활성 함수(sigmoid fuction)로 tangent hyperbolic 함수이다.

다음과 같은 출력 벡터를 갖는 은닉 프레임(j)의 출력은

$$y_j = [y_{j1} \dots y_{jl} \dots y_{jL}]^t \tag{2.2}$$

와 같고 입력 프레임 벡터 $X_j(t)$로부터 유도된 벡터 y_j는 다음과 같이 나타낼 수 있다.

$$y_j = f(\sum_i \beta(j, t) X_j(t)) \tag{2.3}$$

여기서, $\beta(j, t) = [\beta_{ik}(j, t)]$은 계수 행렬이다.

은닉층에서 출력 유닛에 연결된 연결강도 계수가 $\alpha(j, l)$로 주어지면 이 계수에 대한 벡터는 다음과 같다.

$$\alpha_j = [\alpha(j, 1) \dots \alpha(j, l) \dots \alpha(j, L)] \tag{2.4}$$

위와 같은 연결강도 벡터를 사용한 전체 신경망의 출력 Net는 다음과 같다.

$$Net = \sum_j \alpha_j \cdot y_j$$

$$= \sum_j \alpha_j \cdot f(\sum_i \beta(j, t) X_j(t)) \tag{2.5}$$

따라서, 위의 (2.5)식은 아래와 같이 된다.

$$Z = Net \tag{2.6}$$

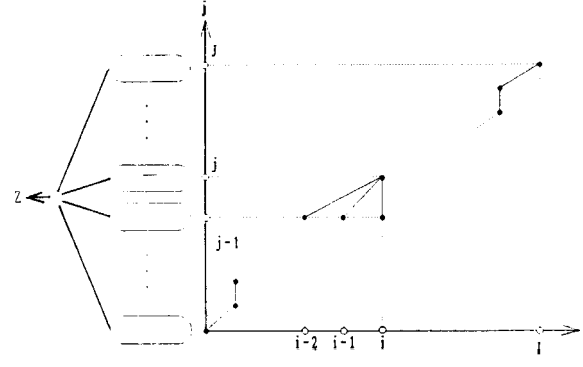

그림 2.2 동적 프로그래밍을 적용한 신경회로망

이와 같은 네트워크은 식(2.5)의 합으로 DP가 연속적인 시간축 정렬의 특징을 가진다. 그리고, 입력층에서의 블럭 구조가 파라메터 패턴 변화의 문제를 해결하기에 적당한 특징을 가진다. 이와 같은 네트워크에 동적 프로그래밍 기술을 적용한다면 그림 2.2와 같은 구조를 보인다.

단순하게 입력 블럭의 길이 τ를 2라고 가정한다. 식 (2.5)에서 입력과 출력과의 관계를 구할 수 있다. 이 수식으로 DP를 사용하여 시간축으로 정렬할 수가 있다.

신경망에 들어갈 입력 패턴이 식 (2.7)과 같다면

$$A = a_1 \cdots a_i \cdots a_I \tag{2.7}$$

그림 2.2에서 이러한 입력 패턴에 대한 시간 정렬이 이루어지는 것을 볼 수가 있다. 그 방법은 다음과 같다. 워핑(Warping)함수 i=i(j)는 입력 패턴의 시간 i와 신경망의 블럭 j와의 관계를 의미한다. 이러한 워핑함수에 의해 신경망의 입력이 결정되는 데 입력 블럭의 길이가 $\tau=2$ 이므로 블럭의 입력은 다음과 같다.

$$X_j(0) = a_{i(j)} \qquad X_j(1) = a_{i(j)-1}$$

즉, 워핑함수에 의해 결정된 $a_{i(j)}$와 시간이 1 만큼 지연된 입력 패턴 벡터 2개가 신경망 블럭의 입력이 된다.

식 (2.5)과 (2.6)에 의해 신경망의 출력은

$$Z = \sum_j \alpha_j \cdot f(\beta(j,0)a_i + \beta(j,1)a_{i-1}) \tag{2.8}$$

이고, 식 (2.8)에서 $\sum$ 안의 식을 $r(i, j)$라고 하면

$$r(i,j) = \alpha_j \cdot f(\beta(j,0)a_i + \beta(j,1)a_{i-1}) \tag{2.9}$$

와 같이 된다. 여기서 시간 정렬은 전체 $r(i, j)$의 합을 최고로 하는 경로에 대해서 이루어진다.

$$Z(A,B) = \max_{i=i(j)} [\sum_j r(i,j)] \tag{2.10}$$

위 식이 바로 입력 패턴 A에 대한 신경망 B의 출력이 $r(i, j)$의 합을 최고로 하는 입력 패턴의 경로를 통해 더해지는 값이라는 것을 의미한다.

2.2 제안된 연속음성에서의 keyword spotting 시스템

본 논문에서 단어를 기본단위로 하는 keyword spotting 시스템은 음소를 기본단위로 하는 keyword spotting 시스템에 비해 단어 내에서의 상호 조음현상(coarticulation)을 별도로 고려할 필요가 없으며, 이로 인해 음소를 기본단위로 하는 방식에 비해 우수한 인식 성능을 얻는 것이 가능하다는 점을 고려하여 DDD 지역명을 대상어휘로 한 247개의 keyword를 선정하였고, 궁극적으로 음성 다이얼링 시스템 구축을 위한 시스템을 제안하였다. 제안된 시스템은 음성 인식을 위해 LPC(Linear Predictive Coding) 분석을 한 후 이로부터 LPC 켑스트럼을 구해 음성 특징으로 사용하며, keyword 인식 과정에서는 keyword 모델과 filler 모델이 병렬 network 구조로 연결된 형태로 이들 사이에 아무런 문법 제약도 가지지 않는 시간동기화를 위한 DP(Dynamic Programming) 정합을 하였다. 여기서, filler model이란 연속음성의 문장에서 non-keyword 부분과 silence부분 들로만 구성된 model이다. 이러한 network 구조를 가지는 전체 시스템의 블럭도는 그림 2.3과 같다.

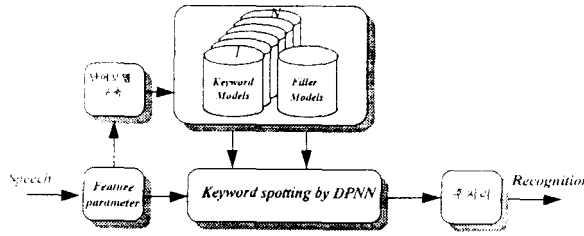

그림 2.3 전체 시스템의 블럭도

본 시스템은 먼저 입력음성이 들어오면 음성신호 전처리 과정에서 음성의 특징 파라메터들을 추출한다. 음성 데이타가 훈련용일 경우 이로부터 단어 DPNN을 훈련시키고, 구성된 단어 모델을 이용하여 keyword 모델 및 filler 모델(non-keyword와 silence)을 구축한다. 그리고, 이들 모델과 keyword spotting과 관련된 문법적 제약 정보(예를 들면, 한 문장 내에 keyword 개수가 한 개 또는 두 개가 들어있다는 등)를 이용하여 전체 DPNN network을 구성한다. 인식 실험용 음성데이타도 동일한 음성 전처리 과정을 거친 다음 구성된 전체 DPNN network 상에서의 DP 정합 과정을 통해 keyword spotting을 수행한다. 이 과정에서 추출된 keyword들의 신뢰도를 판단하기 위해 후처리 과정이 사용되며, 신뢰도가 높은 keyword들만 최종적으로 인식된 keyword로 선정하게 된다.

연속 음성인식에서의 일반적인 keyword spotting 방식은 keyword 모델과 filler 모델을 사용하고 있다. 이 경우

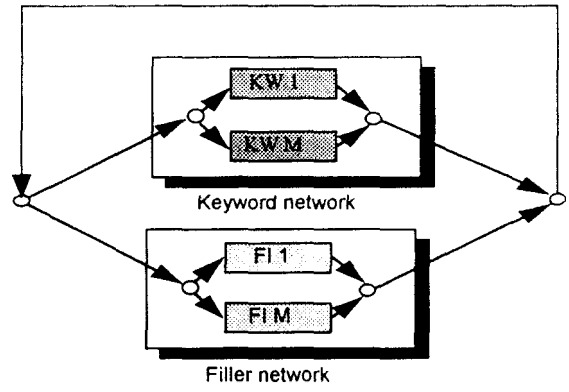

그림 2.4 Null-grammar 형태의 word spotting 모델의 예

filler 모델들은 keyword에 해당되지 않는 음성구간, 즉, non-keyword 구간들과 음성이 아닌 배경잡음 구간들을 성합시키는 데 사용된다. 따라서, 입력된 음성은 이들 keyword들의 시간 순서열로 표현되며, 이 과정에서 keyword가 검출된다. 이와 같은 word spotting 모델의 구성예는 그림 2.4에서 보이고 있다.

본 논문에서는 단어를 모델링하여 전자처럼 non-keyword 각각을 모델링하면 방대한 훈련용 데이터가 필요하므로 후자의 keyword가 아닌 부분 전체를 filler 모델로 구성하였다. Keyword가 아닌 부분 전체를 모델링할 때는 keyword가 아닌 부분을 모아서 단지 몇개의 non-keyword 모델로 표현하면 되므로 데이터량을 많이 줄일 수 있었다. 그밖에 keyword spotting 시스템의 성능향상을 위한 접근방법으로 후처리 과정을 적용하였다. 후처리 과정을 수행하는 목적은 keyword spotting의 성능이 완벽하지 못한 상태에서 실제 응용분야에 적용하기 위해서는 잘못된 결과를 출력시키기 보다는 결과를 출력시키지 않고 재시도를 하는 것이 많은 경우 문제를 줄일 수 있다는 판단에 근거를 둔 것이다[12]. 후처리 방법을 나타내는 구성도는 그림 2.5에 설명되어 있다.

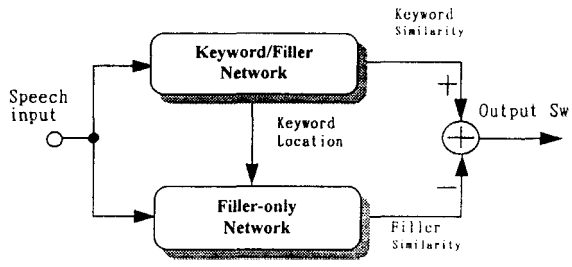

그림 2.5 후처리 방법

Keyword와 filler에 의한 log scoring 방법은 그림 2.5에 나타낸 바와 같이 두가지의 DPNN network을 병렬로 사용한다[4]. 그 중 첫번째는 keyword 및 filler로 구성된 network이고, 두번째는 keyword 모델없이 filler 모델만으로 구성된 network이다. 따라서, keyword 및 filler network는 DPNN에 의한 keyword spotter로서 동작하여 입력 문장으로부터 keyword 후보 및 이 keyword가 존재하는 음성 구간 정보를 검출해 내고, filler만으로 구성된 network는 입력 문장을 filler 모델만으로 표현하므로 앞서 구해진 keyword 후보 위치에 해당하는 filler string을 구할 수 있다. 만일 keyword 및 filler network에 의해 특정 키워드 w가 프레임 ts로부터 프레임 te까지의 구간에서 검출되었다고 하자. 이 때, keyword의 유사도로부터 이 구간에 해당하는 filler 모델의 유사도에 log를 취하여 뺀 값을 Sw라고 하면, Sw는

$$S_w = \log[X_t^t | w] - \log[X_t^t | f] \qquad (2.14)$$

로 주어진다. 여기서, w와 f는 각각 keyword 및 filler 모델을 의미하고, $X_t^t = x_t, \ldots, x_t$는 frame ts로부터 frame te까지의 파라메터 벡터들이다. 이와 같이 구한 Sw를 적절한 경계치와 비교함으로써 keyword 검출여부를 최종적으로 결정한다.

Ⅲ. 인식 실험 및 고찰

3.1 음성 데이터베이스

본 논문에서 실험에 사용된 음성 데이터 베이스는 전국 DDD 지역명 247개의 단어와 각 도명 9개의 단어를 남성화자 6명에 의해서 발음하게 하였다. 이에 대한 keyword들의 목록은 표 3.1에 나타내었다. 표에서 보는 바와 같이 총 인식 대상 keyword의 개수는 247개이다. 먼저 단독어 실험을 위해 4명의 화자가 DDD 지역명을 각 10회 발음한 9,880개의 keyword를 가지고 화자 종속-독립 실험에 사용하였다. 연속음성에서의 keyword spotting 인식 실험을 위해 음성 다이얼링 시스템에서 사용될 수 있는 간

표 3.1 DDD 지역명을 위한 keyword 목록

서울특별시	일 산	철 원	합 덕
부산직할시	의 정 부	동 송	대 천
대구직할시	동 두 천	와 수	부 여
인천직할시	이 천	태 백	서 산
광주직할시	장 호 원	화 광	태 안
대전직할시	전 곡	도 계	해 미
경 기	연 천	평 창	서 천
수 원	평 택	홍 천	한 산
가 평	안 중	화 천	장 항
청 평	청 북	횡 계	예 산
강 화	송 탄	횡 성	삽 교
온 수	원 곡	충 북	온 양
광 주	포 천	청 주	아 산
신 장	강 원	괴 산	둔 포
구 리	춘 천	중 평	조 치 원
김 포	강 릉	단 양	전 의
봉 진	주 문 진	보 은	전 동
문 산	대 화	영 동	천 안
금 촌	동 해	옥 천	성 환
법 원 리	북 평	옴 성	청 양
안 산	삼 척	무 극	홍 성
발 안	근 덕	제 천	광 천
남 양	상 동	진 천	경 북
조 암	속 초	충 주	경 주
오 산	설 악	충 남	건 천
성 남	간 성	신 탄 진	안 강
안 성	거 진	유 성	고 령
안 양	산 동	공 주	구 미
군 포	양 구	유 구	선 산
양 평	양 양	금 산	군 위
용 문	영 월	논 산	김 천
여 주	원 주	연 무 대	봉 화
용 인	인 제	강 경	상 주
고 양	정 선	연 산	성 주
원 당	사 음	당 진	안 동

영 덕	현 풍	창 녕	송 정
영 양	경 남	영 산	강 진
영 주	마 산	남 지	고 흥
풍 기	창 원	충 무	녹 동
영 천	신 북	하 동	곡 성
금 호	거 제	함 안	광 양
예 천	장 승 포	함 양	옥 곡
왜 관	고 현	합 천	구 례
동 명	거 창	전 북	금 성
울 릉	고 성	선 주	담 양
울 진	기 장	삼 례	목 포
후 포	일 광	고 창	무 안
죽 변	양 산	군 산	보 성
평 해	김 해	대 야	벌 교
온 정	진 영	옥 구	순 천
의 성	남 해	김 제	승 주
안 계	지 족	남 원	여 수
탑 리	밀 양	무 주	영 광
점 촌	수 산	부 안	법 성 포
문 경	삼 랑 진	순 창	영 암
가 은	사 천	이 리	완 도
청 도	삼 천 포	함 열	장 성
청 송	산 청	임 실	장 흥
진 보	울 산	장 수	진 도
포 항	언 양	장 계	함 평
오 천	온 산	정 주	해 남
흥 해	의 령	신 태 인	화 순
하 양	진 주	진 안	제 주
구 룡 포	진 해	전 남	

단한 희화체 형식의 서로 다른 문장을 화자 3명이 각 2회씩 발음하도록 하여 총 1,482개의 문장으로 실험하였다.

3.2 특징 파라메터 추출

음성으로부터 특징을 추출하는 방법 중에 하나는 음성 파형 자체를 그 특징으로 생각하는 것이다. 그러나 음성 파형은 시간에 따른 많은 변화가 있고 데이터양도 많아 이를 주파수 영역으로 변환시켜 특징을 추출하는 방법을 사용한다. 다음으로는 음성이 성도로부터 발성된다는 사실에서 구강의 형태를 필터로 가정하여 그 필터 계수를 음성의 특징으로 삼는 것이다. 대개 필터는 AR(Auto Regressive) 혹은 ARMA(Auto Regressive Moving Average) 모델에 의해 구성되는데 AR 모델의 대표적인 것으로 LPC (Linear Predictive Coding) 방식을 사용한다. 또한 귀가 음성을 분석하는 방식을 이용한 청각 특성 분석 방식이 있는데 이는 저주파에서는 상세히, 고주파 영역에서는 개략적으로 측정하여 주파수 영역내의 가중치 함수를 구하여 바크 스케일(Bark scale) 혹은 멜 스케일(Mel scale)이라 명명하였다. 이 스케일에 따라 음성의 특징을 FFT (Fast Fourier Transform)에 의하여 주파수 영역으로 변환시킬 때 가중치를 주는 방법과 LPC에 의해 추출된 파라메터를 가중치로 주어 파라메터로 구하는 방법이 있다.

본 실험에서는 특징 파라메터 추출을 위해 IBM-PC에 부착된 TMS320C30 DSP보드를 이용하여 마이크로 입력된 음성을 70Hz~4.5KHz의 대역 통과 필터링을 한후 10KHz로 샘플링과 16bit 양자화를 통해 A/D 변환하였다. 이 데이터는 256 샘플씩 프레임으로 추출된 후 음성 발생 모델에서 생기는 고주파 감소 현상을 보상하기 위해 프리엠퍼시스(pre-emphasis)필터 $H(z)=1-0.95z^{-1}$를 통과시킨다. 그리고 프레임 추출시 시작 부분과 끝부분에 발생하는 왜곡을 보상하기 위해 해밍창(Hamming Window) $W(n)=0.54-0.46\cos(2\pi n/N-1)$을 씌운다.

3.3 인식 실험

본 연구는 신경회로망을 이용한 연속음성에서의 keyword spotting 실험을 하였다. 본 실험에서는 보다 빠른 학습과 높은 인식률을 내기 위하여 DDD 지역명을 각 도별로 나누어 발음하였다. 데이터는 전국 DDD 지역명 247개의 단어를 남성화자 4명에 의해서 각각 10회씩 발음하게 하였다. 총 인식 대상 keyword의 개수는 247개이며, 전체 데이터베이스의 크기는 9,880개의 단어를 가지고 학습과 단독어 인식실험에 사용하였다.

먼저 DPNN 신경회로망을 구성하여 단어를 학습시킬 때 입력 파라메터를 몇 차로 할것인가? 은닉층의 수를 몇 개로 구성할것인가에 대한 실험을 각각 수행하였다.

표 3.2 파라메터 차수를 달리했을 때의 인식률

차수	10	12	16
경 기	96.4	95.5	93.1
강 원	92.3	97.3	95.5
충 북	98	96.3	89.6
충 남	93.9	91.4	95.7
경 북	91.8	93.6	92.3
경 남	89.6	94.6	93.5
전 북	95.6	95.6	94.5
전 남	96.5	90.1	89.4
전 체	94.26	94.17	93.07

표 3.3 한 프레임당 은닉층의 수에 따른 인식률

은닉층의 수	2	4	8
경 기	92.7	96.4	94.6
강 원	88.6	92.3	90.5
충 북	91.3	98	94.6
충 남	89.1	93.9	92.6
경 북	94.6	91.8	95.1
경 남	90.7	89.6	92.4
전 북	95.6	95.6	96.6
전 남	90.8	96.5	93.7
전 체	91.67	94.26	93.76

위의 실험결과에서 처럼 입력 파라메터를 10차로 했을 때와 은닉층의 수를 4로 했을 때 인식률이 좋았으며, 계산량에 있어서도 뛰어났다.

위의 실험을 기반으로 하여 화자종속-독립에 대한 고립단어 인식 실험을 행하였다. 그 결과는 표 3.4와 3.5에 잘 나타나 있다.

표 3.4 고립단어 인식 (화자 종속)

	화자 1	화자 2
경 기	96.4	93.6
강 원	92.3	94.2
충 북	98	89.6
충 남	93.9	95.1
경 북	91.8	92.7
경 남	89.6	90.2
전 북	95.6	91.4
전 남	96.5	94.4
평 균	94.26	92.65
전 체	93.45	

표 3.5 고립단어 인식 (화자 독립)

	화자 1	화자 2
경 기	85.3	82.1
강 원	86.7	86.1
충 북	79.6	84.6
충 남	83.6	85.4
경 북	84.8	81.6
경 남	81.8	85.7
전 북	82.9	88.2
전 남	85.8	80.8
평 균	83.8	84.30
전 체	84.05	

연속음성에서의 keyword spotting 실험을 위해 간단한 회화체 형식의 서로 다른 문장을 화자 3명이 각 2회씩 발음하도록 하여 총 1,480 문장을 화자종속-독립 실험에 사용하였다.

먼저, 성능향상을 위해 도입하였던 후처리 방법의 타당성을 보이기 위해 후처리를 하지 않았을 때와 후처리 하였을 때의 비교 실험을 하였다.

표 3.6 문장 형태의 인식 실험

	후처리 않함	후처리(○)
문 장(40)	73.8	79.7

표 3.7 문장 형태의 인식 실험 (화자 종속)

	화자 1	화자 2
경 기	75.7	78.5
강 원	76	76.6
충 북	77.6	79.3
충 남	74.4	75.6
경 북	73.9	80.4
경 남	76.5	78.7
전 북	74.4	81.7
전 남	78.8	79.5
평균	75.91	78.78
전 체	77.34	

표 3.8 문장형태의 인식실험 (화자독립)

	화자 1
경 기	72.6
강 원	63.3
충 북	70.6
충 남	74.2
경 북	73.7
경 남	69.2
전 북	72.4
전 남	69.1
전 체	70.63

3.4 고 찰

본 논문에서는 음성신호 전처리 과정에서 기존 음성인식 분야의 연구에서 성능이 우수하다고 판단된 방식을 사용하였고, 이러한 전처리 과정을 통한 음성을 가지고 단어모델을 구축하였다. 그리고 시스템의 성능 향상을 위해 변별력을 갖기 위한 간단한 후처리 과정을 수행한 후 인식을 하였다. 먼저 입력 파라메터의 차수에 따라 인식률과 계산속도에 영향을 미치기 때문에 단독어에 대해 차수를 변화시켜 실험해 본 결과 인식률에는 별 차이가 없었으나 단어를 기본단위로 하는 시스템 전체의 계산속도면에서 차이를 볼 수 있었다. 또한 신경회로망의 분류능력에 대한 실험으로 각 은닉층의 수를 달리하여 실험한 결과 은닉층의 수가 적을 수로 계산속도는 빠르나 분류능력에서 떨어지고 은닉층의 수가 많을수록 분류능력은 뛰어나지만 계산량이 많다는 단점이 있어 이러한 점을 고려하여 실험하였다.

문장 형태의 keyword spotting 실험에서 후처리를 도입하여 약 5%의 성능개선을 보였지만 인식률에서는 저조했다. 이와 같은 결과는 데이터의 부족과 화자마다 발음할때 발음의 길이차에 커다란 영향을 미친 것으로 사료된다. 이와 같은 예를 그림 4.1에 보였다. 그림 4.1(a)는 '서울'이라는 발음이 43 프레임 정도이고 (b)는 54 프레임

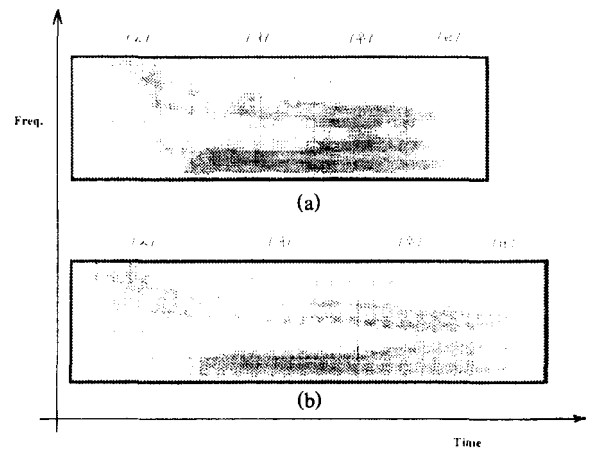

그림 4.1 발음에 대한 지속 시간 차이 (예 : 서울)
(a) 화자 1 (b) 화자 2

의 길이를 보이므로 화자마다 다른 발음의 차이를 볼 수 있다. 발음의 차이는 본 시스템에서 입력된 음성을 제안된 알고리즘으로 어느 정도 흡수할 수 있을 것으로 본다. 그리고 세그멘테이션이 부정확하므로 이를 해결하도록 하였다. 또한 데이터를 발음할 때 정확하게 발음해야 하는데, 그렇지 못한 것이 문제가 되었다. 그러므로 이와같은 문제점을 어느 정도 해결하고 화자의 수와 발음 횟수를 많이 늘려 학습 데이터의 학습이 충분히 이루어진다면 좋은 결과가 나올것으로 기대한다.

Ⅳ. 결 론

본 연구에서는 신경회로망을 이용하여 연속음성에서의 keyword spotting 시스템을 제안하였다. 본 논문에서 다루고 있는 keyword spotting 분야는 사용자가 자연스러운 연속음성으로 말하더라도 이로부터 미리 주어진 핵심 주제어(keyword)들을 검출해 냄으로써 컴퓨터와 인간 사이의 기본적인 의사소통을 가능케 하는 장점 때문에 선진국에서는 최근 활발한 연구가 이루어지고 있는 분야임에도 국내에서는 연구가 별로 이루어지지 못한 실정이었다. 본 논문에서의 인식 성능은 247개의 전체 keyword를 각 도별로 나누어 인식 실험한 결과이다.

제안된 keyword spotting 시스템은 궁극적으로 고립 단어인식에서의 사용자의 불편함과 연속 음성인식의 성능 저조와 같은 일반적인 음성인식 기술의 문제점을 해결하기 위한 기반 구축을 위한 것이다. 적용된 신경 회로망의 장점인 시간축의 왜곡과 스펙트럼의 왜곡을 흡수할 수 있는 DPNN을 이용하여 keyword spotting 실험을 하였다. 총 247개의 keyword를 대상으로 화자종속 고립단어 인식 결과 93.45%의 결과를 보였고, 화자독립 실험 결과 약 84.05%의 인식률을 보였다. 그리고, 연속음성에서의 keyword spotting 실험의 인식 성능 향상을 위해 간단한 후처리 방법을 적용한 결과 40개의 문장을 사용하여 약 5%의 성능 향상이 있었음을 알 수 있었다. 간단한 회화체 형식의 문장 1482개를 대상으로 실험하여 화자 종속은 77.34%의 인식률을 보였고, 화자 독립은 70.63%의 결과를 보였다. 연속어에 대한 실험은 계속 진행중에 있다. 이와 같은 인식 결과로 미루어 볼때 보다 많은 데이터베이스가 구축이 된다면 제안된 keyword spotting 시스템을 대화체 연속 음성인식에 적용하여 보다 새로운 응용분야의 창출에 기여할 수 있을 것으로 기대된다.

Keyword spotting 기술을 대화체 연속 음성인식에 적용하는 방법으로는 전처리기 또는 보조처리기로 사용하는 방안이 검토될 수 있으며, 본 논문에 적용한 방식은 연속 음성인식 시스템의 보조 처리기로 활용될 수 있다. 또한, keyword spotting 기술이 연속 음성인식 시스템의 전처리기로 사용되기 위해서는 앞으로 고속의 keyword spotting 방식이 연구될 필요가 있다.

참 고 문 헌

1. H.Sakoe, R.Isotani, K.Yoshida, "Speaker-Independent Word Recognition Using Dynamic Programming Neural Networks", Proc. ICASSP-89, pp. 29-32, 1989.
2. D.Rumelhart and J.McClelland, "Parallel Distributed Processing", MIT Press, 1986.
3. W.Y.Huang, R.P.Lippmann, and B.Gold, "A Neural Net Approach to Speech Recognition", Proc. ICASSP-88, pp. 99-102, 1988.
4. Mitchel Weintraub, "LVCSR Log-likelihood Ratio Scoring For Keyword Spotting", Proc. ICASSP-95, pp. 297-300, 1995.
5. R. C. Rose and D. B. Paul, "A Hidden Markov model based keyword recognition system," in Proc. IEEE ICASSP, 1990, pp. 129-132.
6. "연속 음성인식에서의 Keyword Spotting 적용방식 연구" 1995. 1. 한국전자통신연구소, 최종연구보고서
7. 이응용 와 4인, "음성 다이얼링 시스템의 구현," 대한전자공학회 추계종합학술대회논집 제16권 제2호, 1993. 11.
8. R. P. Lippmann and B. Gold, "Neural classifiers useful for speech recognition," in 1st International Conference on Neural Network, IEEE June 1987.
9. M. W. Feng and B. Mazor, "Continuous word spotting for application in telecommunications," in Proc. ICSLP, 1992, pp. 619-622.
10. C. Tadj and F. Poirier, "Keyword Spotting using Supervised/Unsupervised Competitive Learning," in Proc. IEEE ICASSP, 1995, pp. 301-304.
11. 한국전자통신연구소, "Word Spotting을 이용한 연속음성 인식 방식연구," 1993년 12월.
12. 김형순, "Keyword Spotting 기술," 한국통신학회지, 제 11권 제 9호, pp. 57-66, 1994년 9월.

▲김 순 협(Soon Hyob Kim): 제14권 5호 참조

▲양 진 우(Jin Woo Yang) 1959년 9월 30일생

1982년 2월: 원광대학교 전자공학과 공학사

1985년 2월: 광운대학교 전자공학과 공학석사

1991년 8월~1994년 8월: 광운대학교 대학원 전자 계산기 공학과 박사 수료

1996년 5월~현재: 춘천 기능 대학 전자 기술학과 전임 강사

관심분야: 음성 인식, 음성 다이얼링, 신경 회로망 등