한국어 변이음 인식을 위한 시간지연 신경망의 확장 방법

°김수일*, 임해창**
* 삼성종합기술원, 기반기술 연구소, 음성처리팀
** 고려대학교, 전산과학과, 자연어처리 연구실

A Method of Scaling Time-Delay Neural Networks for Korean Allophone Recognition

°Soo-il Kim*, Hae-Chang Rim**
* Samsung Advanced Institute of Technology
** Dept. of Computer Science, Korea University

요약

본 논문에서는 한국어 변이음을 인식하기 위한 시간지연 신경망의 확장 방법을 살펴보고 한국어 파열음의 변이음을 인식하는 실험을 통해 각 확장 방법의 인식 성능을 비교한다. 먼저 변이음을 연속음성인식의 인식단위로 사용하기 위하여 한 음소의 모든 변이음을 고려하면서 서로 유사한 변이음을 통합 분류하여 3 개의 변이음 군으로 나눈다. 한국어 파열음에 대한 인식 실험 결과, 음향 음성학적인 특성에 따라 나누어진 소규모 시간지연 신경망들을 모듈 별로 학습한 후, 계층적으로 통합하여 전체적인 시간지연 신경망을 구성하는 방법이 가장 좋은 성능을 나타내었다.

또한, 변이음 단위 인식이 음소 단위 인식에서 문제가 되는 조음결합현상을 해결할 수 있음을 확인하였고, 변이음 인식의 결과인 변이음 열이 제공하는 부가적인 정보를 음운과성에 이용하는 방법에 대해 고찰하였다.

Ⅰ. 서론

음성인식용 신경망은 음성신호의 시간축 왜곡 현상을 처리할 수 있어야 한다. 즉 다음 음절(또는 음소)이 어디서 시작하는 지 정확히 예측할 수 없으므로 해석구간내의 어느 부분에 해당 음절이 위치하더라도 찾아낼 수 있어야 한다. 또한 음성신호의 크기와는 무관한 스케일 불변성(scale-invariant)이 있어야 한다. 이 밖에 음성인식용 신경망은 음성 신호의 포만트(formant) 분포와 같은 정적인 특성과 포만트 천이와 같은 동적인 특성을 추출하는 능력이 있어야 한다. 시간지연 신경망은 이와 같은 불변 특성과 패턴인식 능력을 갖는 대표적인 신경망으로서 음성의 시간에 따른 음향학적 특징 변화를 학습할 수 있기 때문에 시간축 왜곡 현상과 음소간 경계 분리 문제를 자연스럽게 해결할 수 있다.

시간지연 신경망을 이용하여 한국어의 모든 음소를 인식하는 경우 40 여개의 출력노드를 갖는 신경망을 구성해야 한다. 그러나, 40 개 정도의 출력 노드를 갖는 단일구조(monolithic) 시간지연 신경망을 학습시키는 것은 거의 불가능하다. 따라서 분류하는 특성에 따라 나누어진 여러 개의 소규모 시간지연 신경망들을 따로 학습시킨 후 이들을 계층적으로 통합하여 커다란 시간지연 신경망을 구성해야 한다[4].

연속음성인식 시스템의 인식단위를 결정하는 데 고려하는 척도에는 일관성(consistency)과 학습가능성(trainability)이 있다[1][2][3]. 일관성이란 선택된 인식단위는 이것을 포함하는 단어나 문장에서 동일한 특성을 가져야 한다는 것이다. 학습가능성이란 학습하기에 충분한 수의 용례가 존재해야 한다는 것이다. 음소보다 상대적으로 큰 인식단위인 음절, 반음절, 다이폰(diphone) 등은 일관성을 갖지만 모델의 갯수가 너무 많아서 학습하기에 충분한 용례를 찾기 힘들다. 반면에 음소는 그 갯수가 충분히 적어서 학습이 가능하지만 주변 음소들에 영향을 많이 받는 조음결합현상(coarticulation)에 의해 일관성이 없다. 문맥의존(context-dependent) 음운인 트라이폰(triphone)과 변이음(allophone)은 일관성을 갖는 인식단위이지만 갯수가 너무 많아서 학습시키기 힘들다. 그러나 음운간 유사도가 큰 것들을 적절히 통합한 일반화된 트라이폰 또는 변이음(generalized triphones or allophones)으로 음성을 모델링하면 충분히 학습시킬 수 있을 뿐만 아니라 음성인식의 궁극적인 목적인 무제한 어휘의 연속음성을 인식하는 것이 가능하다.

본 논문에서는 음소단위 음성인식에서 문제가 되는 조음결합현상을 극복하기 위해 변이음을 인식 단위로 정한다. 먼저 한 음소의 여러 환경에서 다르게 나타나는 모든 변이음을 고려하고, 유사한 변이음을 통합 분류하여 몇 개의 변이음 군으로 나눈다. 시간지연 신경망을 확장하는 세가지 방법의 인식 성능을 비교하기 위해, 일반적으로 서로 구별하기 어렵다고 알려진 6 가지 한국어 파열음의 변이음에 대해 인식실험을 하였다.

Ⅱ. 변이음 단위 인식

연속음성인식을 위한 기본단위는 명확히 정의되고 학습하기에 충분하도록 자주 발생해야 할 뿐만 아니라 이웃하는 단위에 영향을 받지 않아야 한다. 음소단위 인식에서 문제가 되는 조음결합현상을 해결하기 위해 사용하는 문맥의존(context-dependent) 단위에는 다이폰, 트라이폰, 그리고 변이음 등이 있다. 문맥의존이란 이웃하는 음소와의 관계를 반영한다는 것을 뜻한다. 다이폰과 트라이폰은 모두 모델의 갯수가 너무 많아서 모든 모델을 충분히 학습시킬 수 없다는 문제점이 있다. 따라서 유사한 트라이폰 모델을 통합한 일반화된 트라이폰(generalized triphones)이 제안되어 HMM (hidden Markov model)을 이용한 음성인식에 적용되었으며 상당히 우수한 성능을 보였다. 그러나 트라이폰은 변이음의 많은 발생 요인중 단지 2가지 요인만을 고려한 것으로, 보다 많은 변이 발생 요인을 고려하고 유사한 변이음을 통합하여 구성한 일반화된 변이음(generalized allophones)을 인식단위로 사용하면 기존의 음소를 단위로 한 음성인식의 문제점을 해결할 수 있고 인식률도 높일 수 있다[1][3][5].

본 연구에서는 변이음단위 인식의 성능을 음소단위 인식과 비교하기 위해 한국어 자음중 사로간의 구별이 어렵다고 알려진 6 가지 파열음을 실험대상으로 하였다. 한국어 파열음중 경음을 제외한 평음과 격음 6 가지(ㄱ,ㄷ,ㅂ,ㅋ,ㅌ,ㅍ)를 표 1과 같이 크게 세가지 변이음군으로 나누었다. 이 분류는 변이음의 발생 요인중 단어내 음소의 위치가 변이음간 구별을 가장 명확히 해주는 특성이라고 보고 정한 것이다.

표 1. 변이음 군의 분류 기준

변이음 군	분류 기준	예
제 1 변이음	단어의 첫번째 음절의 초성으로 발음될 경우	(가나)에서 /ㄱ/
제 2 변이음	단어 중간에 있는 음절의 초성으로서 앞 음이 유성음인 경우(이때 유성음은 단모음 8개와 유성자음 8개: ㅏ,ㅓ,ㅗ,ㅜ,ㅡ,ㅣ,ㅐ,ㅔ,ㄴ,ㄹ,ㅁ,ㅇ)	(아가)에서 /ㄱ/ (창가)에서 /ㄱ/
제 3 변이음	단어의 마지막 음절의 종성	(사적)에서 /ㄱ/

표 1의 분류에 포함되지 않는 변이음은 다음과 같은 규칙에 의해 처리된다.

(1) 연음 규칙: '북어'는 발음상 /부거/가 되어 첫째 음절의 종성 'ㄱ'은 연음되어 제 2 변이음으로 분류하는 것이 적합하다. 연음으로 발음된 단어를 원래의 단어로 인식하려면, 이 단어가 발음될 때의 문맥 정보를 고려해야 한다.

(2) 경음화: '숙적'은 발음상 /숙쩍/이 되어 둘째 음절의 초성 'ㅈ'은 /ㅉ/으로 경음화되었다. 경음화된 단어를 원래의 단어로 인식하려면, 먼저 경음 인식기를 추가하여 경음을 인식하고 인식된 경음을 문맥 정보에 따라 원래의 음으로 변환해야 한다.

(3) 기타 여러 음운 규칙에 의해 생성되는 변이음은 위에서 분류한 군 중의 하나로 인식한 후, 문법이나 문맥 정보를 이용한 후처리 부분에서 원래의 음으로 변환한다.

변이음단위 인식실험 이전에 행한 파열음소의 각 변이음간 분류 실험을 통해 이와 같은 분류가 타당한지 알아 보았다. 아래 표 2의 실험 결과에서 보면, 각 파열음소의 변이음간 구별이 'ㄱ'의 변이음들을 제외하고는 모두 높은 인식률과 함께 인식됨을 알 수 있다. 이것은 파열음의 각 변이음은 음향 음성학적으로 명확히 구별되는 특성을 갖는다는 것을 의미하고, 이 변이음을 한가지 음소로 일괄적으로 인식하는 것이 인식률 저하의 요인이 됨을 알 수 있다.

표 2 파열음소의 변이음간 분류 실험 결과

TDNN 종류	학습률(%)	인식률(%)	인식오류수 / 입력데이터 수
ㄱ1/ㄱ2/ㄱ3	100	100	0 / 75
ㄷ1/ㄷ2/ㄷ3	100	96.00	3 / 75
ㅂ1/ㅂ2/ㅂ3	100	98.67	1 / 75
ㅋ1/ㅋ2	100	86.00	7 / 50
ㅌ1/ㅌ2	100	100	0 / 50
ㅍ1/ㅍ2	100	98.00	1 / 50

III 시간지연 신경망을 이용한 음성인식

1. 시간지연 신경망의 구조와 학습 방법[6][7][8]

그림 1은 'ㄱ'의 세가지 변이음(/k/, /g/, /k'/)을 구별하는 시간지연 신경망의 구조를 나타낸 것이다. 이것은 출력노드가 3 개인 3 계층 시간지연 신경망이다. 여기에서 입력 패턴의 세로축은 16 구간의 멜 스케일로 변환된 주파수 구간을 나타낸다.

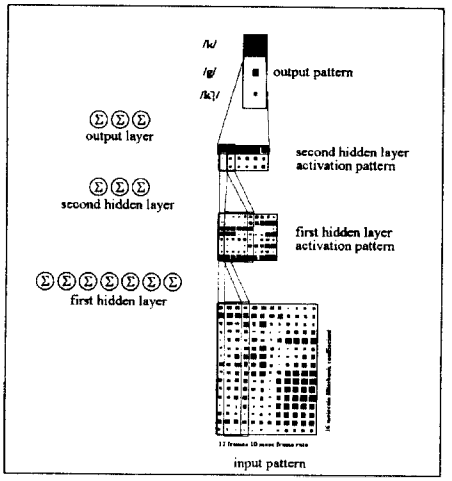

그림 1 시간지연 신경망의 구조

전체적인 인식과정은 다음의 순서대로 이루어진다. 첫번째 은닉층에서는 3개의 지연된 입력을 동시에 받아들여서 가중치(weight)를 곱한 후 시그모이드(sigmoid) 함수를 거쳐 두번째 은닉층에 전달한다. 두번째 은닉층에서는 첫번째 은닉층에서 생성된 활성화 패턴으로부터 5 개의 지연된 입력을 동시에 받아들여서 가중치를 곱한 후 시그모이드 함수를 거쳐 출력층에 전달한다. 출력층에서는 6 개의 지연된 입력을 동시에 받아들인다. 그러나 은닉층에서와는 달리 가중치를 곱하지 않고 단순히 각 출력 노드에 해당하는 두번째 은닉층 노드의 모든 활성화 패턴을 합하여 시그모이드 함수를 취한다. 즉 출력층의 역할은 단순히 두번째 은닉층의 결과를 합하여 최대값을 갖는 출력 노드를 활성화시키는 것이다.

시간지연 신경망의 학습은 오류역전파(error back-propagation) 알고리즘으로 이루어진다. 그러나 시간적으로 지연된 입력 윈도우가 중복되어 입력되기 때문에 다음과 같은 변경이 필요하다.

1) 일반적인 다계층 신경망에서 첫번째 은닉층의 오류는 다음 공식을 이용하여 두번째 은닉층으로부터 전달된다.

$$\delta_i = x_i(1-x_i)\sum_j \delta_j \omega_{ij}$$

여기서 δ_j와 δ_i는 각각 첫번째 은닉층 노드 j의 오류, 두번째 은닉층 노드 i의 오류를 의미한다. x_i는 노드 j의 활성화 정도이고, ω_{ij}는 두번째 은닉층 노드 i에서 첫번째 은닉층 노드 j로 연결된 가중치이다. 시간지연 신경망에서는 시간지연된 값이 동시에 입력되고, 전체적으로 입력 윈도우가 10 ms 간격으로 이동하면서 중복하여 입력되기 때문에 첫번째 은닉층의 활성화 패턴은 여러 개의 두번째 은닉층의 활성화 패턴을 생성하는데 중복해서 사용된다. 따라서 순방향 진행시(forward pass) 연결되었던 모든 경우에 대하여 오류를 계산하고 이들의 합을 실제 역전파된 오류로 사용한다. 즉 첫번째 은닉층의 오류는 다음과 같은 공식을 이용하여 두번째 은닉층으로부터 전달된다.

$$\delta_{ij} = x_{ij}(1-x_{ij})\sum_{t2}\sum_{i}\delta_{it}\omega_{ij}$$

여기서 δ_{ij}는 첫번째 은닉층의 노드 j가 t_1프레임에서 입력한 활성화 패턴 x_{ij}에 대하여 두번째 은닉층으로부터 전달된 오류이다. δ_{it}는 두번째 은닉층의 노드 i가 t_2 프레임의 활성화 패턴을 생성할 때 발생한 오류이다.

2) 입력 윈도우가 10 ms 간격으로 이동하면서 처리하기 때문에 같은 가중치를 사용하여 입력 패턴을 입력하게 된다. 따라서 오류 역전파시 각 윈도우에 대한 오류의 평균을 실제 발생한 오류로 사용한다.

2. 계층구조 시간지연 신경망[4]

한국어의 모든 음소를 인식하는 경우 40 개의 출력 노드를 갖는 시간지연 신경망이 필요하게 된다. 그러나 40 개의 출력 노드를 갖는 단일구조(monolithic) 시간지연 신경망을 학습시키는 것은 거의 불가능하다. 따라서 여러 개의 소규모 시간지연 신경망을 학습시킨 후 이들을 통합하여 커다란 시간지연 신경망을 구성하는 방법이 필요하다.

계층구조 시간지연 신경망은 음소군 분류(phoneme group spotting) 시간지연 신경망과 음소 분류(phoneme spotting) 시간지연 신경망을 모듈별로 학습시킨 후, 음소군 분류 시간지연 신경망을 상위계층으로 하여 각 음소 분류 시간지연 신경망을 계층적(hierarchical)으로 연결함으로써 전체 음소 인식 시스템으로 확장하는 것이다.

계층 구조 시간지연 신경망에서는 단일구조 시간지연 신경망의 두번째 은닉층을 학습하는 과정이 생략되지만 단일구조 시간지연 신경망과 동등한 인식률을 얻을 수 있다.

Ⅳ. 실험 방법

1. 변이음 추출을 위한 음성 데이타 생성

본 연구에서는 제안한 변이음 분류 방법에 의해 6 개의 한국어 파열음을 유성 파열음은 각각 3 가지 군으로 무성 파열음은 각각 2 가지 군으로 나누어 모두 15 가지 군으로 구분하였다. 이때 무성 파열음인 'ㅋ', 'ㅌ', 'ㅍ' 는 종성의 대표음화 현상에 의해 제 3 변이음을 구분할 필요가 없다. 각 변이음의 추출을 위한 단어의 선정 방법은 다음과 같다.

(1) 제 1 변이음: 해당 파열음을 초성으로 하고 8 가지 단모음과 결합되는 단어들을 선정하였다. 이때 음의 변이에 거의 영향을 끼치지 않는 종성 결합 여부는 고려치 않았다.

(2) 제 2 변이음: 단모음 8 개와 유성자음 4 개로 구성된 유성음들 사이에 해당 파열음이 있는 단어를 선정하였다. 그러나 '산길'/산낄/ 과 같이 경음화되는 단어는 제외하였다.

(3) 제 3 변이음: 단어의 마지막 음절 종성으로 해당 파열음이 나타나는 단어를 선정하였다.

각 군마다 해당 변이음을 포함하는 임의의 단어를 25 개씩 선택한 후 한 사람의 남성 화자가 한 단어를 2 번씩 발음하여 녹음하였다. 실험에 사용된 음성 데이타의 수는 각 변이음 마다 50 개씩이고 모두 합하여 750 개이다. 녹음 방법은 컴퓨터가 작동하고 있는 일반 사무실 환경에서 발음한 음성을 마이크를 통해 녹음하고, 16 kHz로 샘플링(sampling)하여 14-bit 데이타열로 변환한 다음 PC에 저장하였다.

2. 음성 신호 처리

시간지연 신경망의 입력 패턴을 생성하기 위한 음성 신호 처리 과정은 그림 2와 같다.

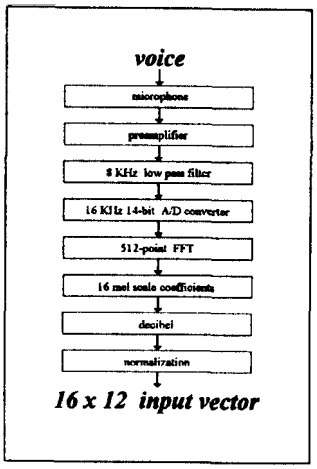

그림 2. 음성 신호 처리

음성 데이타의 스펙트로그램(spectrogram)을 통해 모음 시작부분을 중심으로 142 ms의 변이음을 추출하였다. 추출한 음소를 512-포인트 해밍 윈도우(Hamming window)를 10 ms씩 오른쪽으로 옮겨 가면서 다시 512-포인트 FFT(fast Fourier transform)를 수행하여 256x12의 파워 스펙트럼 계수(power spectrum coefficients)를 만들었다.

낮은 주파수의 음성 신호에 더 민감한 귀의 특성을 반영하기 위해 파워 스펙트럼 계수를 다시 16x12의 멜 스케일 필터뱅크(mel-scale filterbank) 계수로 변환하였다[9]. 그리고 소리의 크기에 대한 귀의 특성을 고려하기 위해 멜 스케일 필터뱅크 파워 스펙트럼을 다시 데시벨(decibel)로 변환하였다[10].

3. 시간지연 신경망의 학습 및 실험

음소 및 변이음 인식을 위한 시간지연 신경망의 노드 수는 아래의 표 3과 같으며, 이 노드 수는 실험을 통해 인식률이 가장 좋은 경우일 때의 값이다.

표 3. 시간지연 신경망의 노드수

TDNN 종류	1st Hidden Layer	2nd Hidden Layer	Output
ㄱ/ㄷ/ㅂ/ㅋ/ㅌ/ㅍ	4	2	2
단어초/유성음사이/단어말 ㄱ/ㄷ/ㅂ	7	3	3
ㅋ/ㅌ/ㅍ ㄱ1/ㄷ1/ㅂ1 ㄱ2/ㄷ2/ㅂ2 ㄱ3/ㄷ3/ㅂ3 ㅋ1/ㅌ1/ㅍ1 ㅋ2/ㅌ2/ㅍ2	9	3	3

시간지연 신경망은 다계층 신경망에 비하여 가중치의 갯수는 적지만 입력 윈도우가 10 ms 간격으로 이동하면서 반복적으로 적용되기 때문에 실제적인 계산량은 다계층 신경망에 비하여 많다. 이에 따라서 학습시간도 급격히 증가한다. 본 연구에서 학습속도를 개선하는데 사용된 방법은 다음과 같다[8][11][12].

(1) 오류 역전파시에는 첫번째 은닉층의 오류를 계산할 때 순방향 진행시(forward pass) 연결되었던 모든 경우에 대하여 오류를 계산하여야 한다. 이에 두번째 은닉층에서 전달된 오류의 평균이 아닌 합을 사용하여 오류계산에 드는 비용을 줄이고 학습시간을 줄였다.

(2) 두번째 은닉층의 활성화된 값을 재곱하지 않고 단순히 합하여 출력층에 전달하였다.

(3) 무작위로 입력 패턴을 배열하고 모든 입력 패턴을 적용한 후 가중치 조정 간격을 점차 늘려갔다.

(4) 시간지연 신경망에 입력되는 입력벡터의 각 에너지값을 -1.0~1.0이 되도록 정규화하였다.

(5) 에러가 E_{MIN}이하인 입력 패턴에 대해서는 학습을 생략하였다 (본 연구에서 사용한 E_{MIN} 값은 0.001이다.)

(6) 학습률(learning rate) η는 학습이 진행됨에 따라서 0.01~0.1사이에서 동적으로 변경된다.

(7) 모멘텀(momentum) α는 학습이 진행됨에 따라서 0.1~0.9 사이에서 동적으로 변경된다.

V. 실험 결과

1 인식 결과

그림 3과 그림 4는 각각 기존의 음소인식 계층구조 TDNN과 본 논문에서 제안한 변이음단위 인식을 위해 구축한 계층구조 TDNN을 나타낸다.

두 계층구조 TDNN에서 사용된 음소군 분류 TDNN, 음소 인식 TDNN, 변이음군 분류 TDNN, 그리고 변이음 인식 TDNN의 인식률은 표 4와 표 5에 나타내었다.

생성된 변이음 데이타 750 개중 절반은 학습에 사용하였으며, 학습에 사용되지 않은 나머지 절반의 데이타를 인식 실험을 위해 사용하였다. 'ㄱ', 'ㄷ', 'ㅂ', 'ㅋ', 'ㅌ', 'ㅍ' 을 분류해내는 음소인식 계층구조 시간지연 신경망을 그림 3과 같이 구축하고 인식 실험한 결과 전체 파열음에 대해 81.07%의 인식률[1]을 얻었다.

본 연구에서 제안한 변이음단위 인식을 위하여 구성한 계층구조 시간지연 신경망을 그림 4와 같이 구축하고 인식 실험한 결과 전체 파열음에 대해 79.07%의 인식률[2]을 얻었다.

두 계층구조 TDNN에서 모두 사용된 유성음과 무성음을 분류 하는 ㄱ/ㄷ/ㅂ/ㅋ/ㅌ/ㅍ TDNN에서 전체 인식률을 저하시키는 대부분의 인식 오류가 발생함을 알 수 있다.

[1] 동일 데이타에 대한 오류가 상위 인식기와 하위 인식기에서 중복되어 발생하면 한 개의 오류로 간주
[2] (기와 동일).

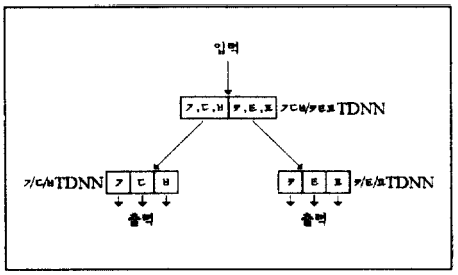

그림 3. 음소인식을 위한 계층구조 TDNN

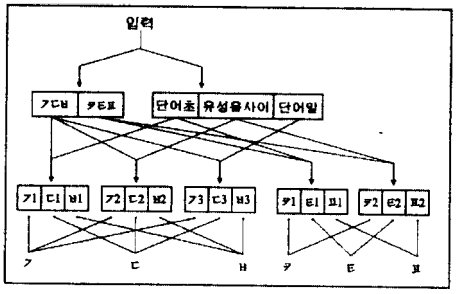

그림 4. 변이음 인식을 위한 계층구조 TDNN

표 4. 음소 인식 오류 갯수와 음소 인식률

TDNN 종류	인식오류 갯수 / 입력데이타 갯수	인식률(%)
ㄱㄷㅂ/ㅋㅌㅍ	48 / 375	87.20
ㄱ/ㄷ/ㅂ	19 / 225	91.56
ㅋ/ㅌ/ㅍ	13 / 150	91.33
합계	71 / 375	81.07

표 5 변이음 인식 오류 갯수와 변이음 인식률

TDNN 종류	인식오류 갯수 / 입력데이타 갯수	인식률(%)
단어초/유성음사이/단어말	21 / 375	94.40
ㄱㄷㅂ/ㅋㅌㅍ	48 / 375	87.20
ㄱ1 / ㄷ1 / ㅂ1	7 / 75	90.67
ㄱ2 / ㄷ2 / ㅂ2	4 / 75	94.67
ㄱ3 / ㄷ3 / ㅂ3	2 / 75	97.33
ㅋ1 / ㅌ1 / ㅍ1	6 / 75	92.00
ㅋ2 / ㅌ2 / ㅍ2	5 / 75	93.33
합계	78 / 375	79.20

그 이유를 살펴보면, 'ㄱ'의 변이음인 /k/(예: '갈')과 'ㅋ'의 변이음 /k^h/(예: '칼')과 같이 동일 화자가 발음하더라도 스펙트로그램 등의 음향 음성학적 특징이 거의 구별할 수 없을 만큼 유사하여 유성음/무성음 분류기 (ㄱㄷㅂ/ㅋㅌㅍ TDNN)에서 초성 파열음이 오인식되는 경우가 많기 때문이다. 이와 같은 현상은 음소단위 인식과 변이음단위 인식에서 모두 발생되는 문제로서 상위 인식기의 오류를 근본적으로 줄여야 해결될 수 있다.

전체 인식률에서는 음소단위 인식이 변이음 인식보다 약간 더 나은 성능을 보였지만, 하위 인식기에서의 성능은 오히려 변이음단위 인식이 우수함을 알 수 있다. 즉 변이음단위 인식 TDNN의 5 개 하위 인식기에서 발생한 오류 갯수는 24 개로서 음소단위 인식 TDNN의 두 하위 인식기에서 발생한 오류 갯수 32 개 보다 적다. 이것은 음소단위 인식의 문제점인 조음결합현상에 의해 발생했던 오류가 변이음단위 인식에서는 발생하지 않았기 때문이다.

그림 5와 그림 6은 변이음단위 인식실험에 사용했던 또 다른 계층구조 시간지연 신경망들이다. 그러나 이 두 구조에 의한 실험에서는 하위 계층 시간지연 신경망의 노드 수가 많아서 학습이 잘 이루어지지 않는 문제점을 나타내었다. 이와 같은 계층구조 시간지연 신경망을 학습시키기 위해 점차적으로 학습 데이타의 갯수를 증가시키는 방법을 사용해보았다[6]. 그 결과, 모든 학습 데이타를 한꺼번에 학습시키는 경우보다 학습률이 향상되는 것을 확인하였으나 그림 4의 계층구조 시간지연 신경망을 사용하는 것만큼 인식률이 향상되지 않았다.

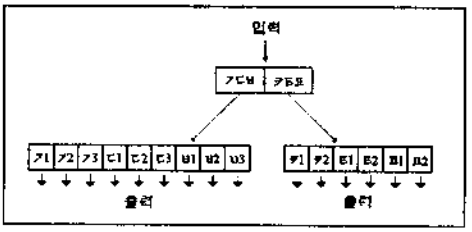

그림 5. 변이음 인식을 위한 계층구조시간지연 신경망 2

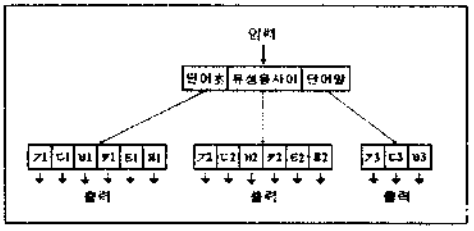

그림 6. 변이음 인식을 위한 계층구조 시간지연 신경망 3

변이음단위 인식의 성능을 보다 향상시키려면, 음소의 단어내 위치 뿐만 아니라 운율, 억양, 강세와 같은 보다 많은 변이 요소를 고려하고 이것을 분류하는 인식기를 추가하는 것이 필요하다.

2. 변이음 인식결과를 이용한 음성인식기의 후처리 방법[13]

'아버지가방에들어가신다.'라는 문장이 입력되었을 때, '가방'의 /ㄱ/과 /ㅂ/이 위의 변이음 군 중 제 1 변이음 또는 제 2 변이음으로 적절히 분류된다면, 구문적인 파싱(parsing)이 가능해진다. 먼저, /ㄱ/이 제 2 변이음, /ㅂ/이 제 1 변이음으로 인식되었다면, 위 문장은 '아버지가방에 들어가신다.'라는 명확한 의미를 가진 문장으로

해석될 수 있다. 반면에, /ㄱ/이 제 1 변이음, /ㅂ/이 제 2 변이음으로 인식되었다면, 위 문장은 '아버지 가방에 들어가신다'라는 문장으로 분석되고 위의 첫번째 문장과는 전혀 다른 의미로 해석될 것이다. 이처럼 변이음 인식 결과는 음소인식에서는 기대할 수 없는 유익한 정보까지도 출력하므로 부가적인 잇점이 있다.

VI. 결론 및 추후 연구 과제

본 논문에서는 한국어 변이음을 인식하기 위한 시간지연 신경망의 확장 방법에 대하여 논하였다. 한국어 파열음의 변이음을 인식하는 실험에서, 소규모 시간지연 신경망을 계층구조로 묶어서 전체 한국어 인식 시스템으로 확장하는 것이 가장 좋은 인식 성능을 보였다. 또한 연속음성인식에서 음소를 기본 인식단위로 할때 발생하는 조음결합현상을 극복하는 방법으로서 변이음단위로 인식하는 것이 유리함을 알 수 있었다.

변이음단위 인식과 음소단위 인식의 성능을 비교하기 위해 한국어의 6 가지 파열음에 대한 인식실험을 하였으며, 각각 79.20%와 81.07%의 성능을 나타내었다. 이것은 상위 계층의 오류가 그대로 하위 계층 오류되어 발생된 결과이다. 그러나, 하위 인식기들의 성능만을 비교하면 변이음단위 인식이 음소단위 인식보다 적은 오류를 나타냄을 확인하였다. 이와 같은 결과는 음소단위인식에서 문제가 되는 조음결합현상이 변이음단위 인식에서는 나타나지 않았음을 나타낸다. 또한 변이음단위 인식에서 출력 결과로 나오는 변이음열은 음운 파싱에 이용될 수 있는 정보를 제공해 주는 부가적인 잇점이 있다.

계층구조 시간지연 신경망의 상위 계층 오류를 하위 계층에서 검출해 내는 방법을 찾는다면 음소 또는 변이음 인식 계층구조 시간지연 신경망의 성능 향상이 가능하다. 또한 운율, 억양, 강세 등의 보다 많은 변이요소를 고려한다면 인식성능 향상과 함께 출력된 변이음 열이 갖고 있는 유익한 정보를 이용할 수 있다.

참고문헌

[1] Kai-Fu Lee, "Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech Recognition", IEEE Trans. ASSP, vol.38, no. 4, pp. 599-609, April 1990.

[2] Hsiao-Wuen Hon, Kai-Fu Lee, "On Vocabulary-Independent Speech Modeling", Proceedings of ICASSP, pp 725-728, April 1990.

[3] Kai-Fu Lee et al., "Allophone Clustering for Continuous Speech Recognition", Proceedings of ICASSP, pp. 749-752, April 1990.

[4] Hidefumi Sawai et al., "Parallelism, Hierarchy, Scaling in Time-Delay Neural Networks for Spotting Japanese Phonemes/CV-Syllables", Proceedings of IEEE International Joint Conference on Neural Networks, vol. 2, pp. 81-88, June 1989.

[5] Chin-Hui Lee, Lawrence R. Rabiner, "Automatic Speech Recognition - Current State and Future Directions", Proceeding of ATR International Workshop on Speech Translation, 1993.

[6] Alex Waibel et al., "Phoneme Recognition Using Time-Delay Neural Networks", IEEE Trans. ASSP, vol. 37, no. 3, pp. 328-339, March 1989.

[7] Richard P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magazine, pp. 4-22, April 1987.

[8] D. E. Rumelhart et al., "Learning Internal Representations by Error Propagation" in D. E. Rumelhart et al., Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1, pp. 318-362, MIT Press, 1986.

[9] S. Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, p. 32, 1991.

[10] D.O'Shaughnessy, Speech Communication, Addison-Wesley Publishing Company, p. 141, 1990.

[11] P. Haffner et al., "Fast Back-Propagation Learning Methods for Neural Networks in Speech", Technical Report TR-I-0058, ATR Interpreting Telephone Research Laboratories, November 1988

[12] Kevin J. Lang et al., "A Time-Delay Neural Network Architecture for Isolated Word Recognition", Neural Networks, vol. 3, pp. 23-43, 1990.

[13] 김가호, "음성인식과 합성에 있어서의 음성학과 음운론의 역할", 제1회 음성학 학술대회 자료집, pp. 125-144, 1994년 2월.