論文96-33B-8-12

KL 변환을 이용한 multilayer perceptron에 의한 한국어 연속 숫자음 인식

(Korean continuous digit speech recognition by multilayer perceptron using KL transformation)

朴 廷 繕 **, 權 章 禹 *, 權 五 常 *, 李 應 赫 ***, 洪 勝 弘 *

(Jungsun Park, Jangwoo Kwon, Eunghyuk Lee, and Seunghong Hong)

요 약

본 논문에서는 신경회로망인 MLP를 이용하여 한국어 숫자음 인식을 시도하였다. 동적인 신호의 인식에서 시간적 상관성을 잘 반영하지 못하기 때문에 잘 사용되지 않는 MLP를 본 논문에서 사용한 이유는 한국어 음절에서는 한국어 음절의 특징을 이용하여 한 음절로부터 정적인 특징을 추출할 수 있었기 때문이었다. 또한 MLP의 전처리 단계로 Karhunen-Loeve변환을 제시하여 사용하였는데 이 변환은 입력 신호를 압축시키면서 각각의 신호의 특징을 잘 구별할 수 있게 하는 특징이 있지만 변환을 통해서 물리적인 특성을 잃게 된다는 단점도 있다. 본 논문에서 음절의 길이에 영향을 덜 받으면서도 정적인 특징을 추출하는 방법을 제시하였기 때문에 KL변환을 MLP의 전처리 단계로 효과적으로 사용할 수 있었다. KL변환을 사용함으로써 인식률을 희생시키지 않으면서도 학습 시간 및 사용 메모리를 감소시킬 수 있었다. 본 논문에서 제안한 음절의 특징 추출은 한국어 음절의 특징을 이용하여 음절의 앞 뒤 일정한 두 구간의 정적인 구간에서 특징을 추출하는 방법인데 이 방법은 일정한 크기의 윈도우를 사용하여 프레임을 만든다는 장점이 있으며 이 방법을 사용하여 기존의 신경회로망에서는 적용하기 힘든 음절의 구분이 모호한 연속음에도 적용할 수 있었다.

Abstract

In this paper, a new Korean digit speech recognition technique was proposed using Multilayer Perceptron(MLP). In spite of its weakness in dynamic signal recognition, MLP was adapted for this model, because Korean syllable could give static features. It is so simple in its structure and fast in its computing that MLP was used to the suggested system. MLP's input vectors was transformed using Karhunen-Loeve Transformation(KLT),which compress signal successfully without losing its separateness, but its physical properties is changed. Because the suggested technique could extract static features while it is not affected from the changes of syllable lengths, it is effectively useful for Korean numeric recognition system. Without decreasing classification rates, we can save the time and memory size for computation using KLT. The proposed feature extraction technique extracts same size of features from the tow same parts,front and end of a syllable. This technique makes frames,where features are extracted,using unique size of windows. It could be applied for continuous speech recognition that was not easy for the normal neural network recognition system.

* 正會員, 仁荷大學校 電子工學科
(Dept. of Electronic Eng., Inha Univ.)
** 正會員, (주) 픽셀 시스템
(PIXEL SYSTEM Ltd.)
*** 正會員, 建陽大學校 컴퓨터 工學科
(Dept. of Computer Eng., Keonyang Univ.)
接受日字:1996年1月22日, 수정완료일:1996年6月24日

Ⅰ. 서 론

본 논문에서는 음성 인식에 사용되는 기존의 신경회로망 중 MLP(Multilayer Perceptron)를 이용하여 MLP의 장점을 살리고 단점을 보완하여 음성신호 중 한국어 숫자음의 인식을 시도해 보았다.

MLP는 특성인 대규모 병렬 처리 기능을 이용하여 하드웨어 구현 시 실시간 처리가 용이하고 적응 학습 기능을 통해 패턴의 결함을 극복하는 능력을 가진다. MLP등은 Rumelhart 등에 의해 역전파 학습 방법이 소개되어 여러 분야에 많이 적용이 되어 왔지만 MLP의 경우 인식할 입력 패턴의 수가 많게 되면 학습 시간이 엄청나게 늘어나게 되는 단점을 가지게 된다. 또한 음성신호의 경우 모음의 포만트와 같은 정적인 특성과 음소 및 단어의 결합 과정에서 나타나는 조음 현상이나 발음 길이에 따른 동적인 특성을 동시에 갖고 있다. 이러한 음성신호에서 정적인 특성을 잘 추출해 낼 수 있다면 MLP는 음성 인식기로서 매우 유용한 방법이 될 수 있다. 따라서 본 논문에서는 음성에서 비교적 정적인 특징을 추출하기 용이한 음절 단위의 인식을 시도하였다.

본 논문에서는 MLP에 전처리 과정을 둠으로써 앞서 말한 MLP의 단점인 학습 시간이 많이 걸린다는 단점을 해결할 수 있었다.

전 처리로써는 KL(Karhunen-Loeve) 변환을 사용하였는데 이러한 KL변환을 MLP의 전처리로서 사용할 때 다음과 같은 장점을 가질 수 있다.

첫째로, KL 변환을 사용함으로써 MLP로 학습될 입력 벡터의 크기를 줄일 수 있었다. 학습 벡터의 크기가 감소함으로써 전체적인 학습 시간 감소 및 가중치 벡터의 크기가 줄어들기 때문에 사용 메모리 감소의 효과를 얻을 수 있었다.

둘째로 KL변환은 변환을 통해서 인식할 패턴 사이의 상호 분산을 0으로 만들어 준다. MLP의 특성상 MLP의 가중치 벡터의 어느 한 요소의 변화는 모든 패턴의 변화에 크거나 작은 영향을 끼치게 된다. 이때 입력 벡터의 상호 분산이 0이 됨으로써 서로의 상관성을 최소화시켜 학습의 빠른 수렴이 이루어지게 된다는 장점이 있다.

또한 본 논문에서는 한국어 숫자음 인식을 할 때 한국어 음절의 특성을 이용하여 음절 길이에 영향을 적게 받는 음절의 특징을 추출 하였다. 한국어의 음성은 음절 단위로 표시할 수 있고 음절은 초성과 중성 그리고 종성으로 이루어져 있는데 모음이 아닌 초성과 종성의 음성신호의 길이는 모음인 중성에 비해서 짧고 일정하며 음절의 전체 길이는 모음의 길이에 가장 많이 영향을 받는다는 것에 착안하여 음성 신호의 시작점과 끝점 구간에서 일정 구간을 취해 MLP의 학습 벡터로 사용하였다.

제안한 모델의 유용성을 입증하기 위해 3명의 남녀 화자가 발음한 고립 숫자음에 대해 인식 실험을 수행하고 1명의 남성 화자가 발음한 2자리 연속 숫자음에 대해서도 인식 실험을 수행하였다.

본 논문에서는 KL변환의 전처리 과정 및 한국어 음절의 특징을 이용하여 기존 MLP장점을 살리고 단점을 보완한 음성 인식 모델을 구성하였으며 또한 이 모델이 연속 음 인식에도 적용될 수 있음을 보였다.

Ⅱ. MLP

MLP는 정적인 패턴 인식에 좋은 성능을 보이기 때문에 음성 인식에 적용할 경우 음절 단위의 인식을 하는데 적당하다. 특히 한국어 음성에는 이중 자음이 없고 초성, 중성, 종성으로 구성되는 음절 단위로 잘 구분되는 특성이 있기 때문에 MLP를 적용하기가 쉽다고 볼 수 있다. 하지만 MLP는 TDNN(Time Delayed Neural Network)와 같은 시간에 따른 음소의 상관성을 고려하지 않았기 때문에 여러 가지 해결해야 할 문제가 있다.

MLP로 음절을 인식하는 경우에 보통 다음과 같은 조건이 요구된다.

■ 일정한 크기의 입력 벡터

일정한 크기의 입력 벡터를 만들기 위해서는 일정한 수의 프레임을 만들어야 한다. 음절의 길이가 발음할 때마다 일정치 않기 때문에 음절을 일정한 프레임으로 나누기 위해서는 프레임을 만드는 윈도우의 크기를 각 음절 마다 적절히 조정해야 한다.

■ 연속 음 적용 시 음절의 명확한 구분

연속 음의 각 음절마다 일정한 수의 프레임을 만들기 위해서는 각 음절의 구분이 명확해야 한다.

가변 윈도우를 사용하여 프레임을 만들 경우 실시간 처리를 위해 하드웨어로 구현하기가 더 힘들어 질 뿐만 아니라 일단 음절이 끝을 검출한 후에 프레임을 만들어 음성의 특징 파라메타를 추출하게 된다. 따라서

인식 시에 시간이 더 소요된다. 또한 음절의 구분이 모호한 연속 음에는 적용할 수 없게 된다는 문제점이 야기된다.

본 논문에서는 제안한 방법은 한국어 음절의 특징을 이용하여 이러한 문제점을 해결하고 MLP에 사용되는 메모리 크기 및 학습 속도를 줄이기 위해서 입력 벡터의 크기 및 상관성을 줄여 주는 전처리 과정을 도입하였다.

Ⅲ. 전처리 및 인식 기법

1. 특징 벡터 추출

본 연구에서 제안한 한국어 숫자음의 특징 추출 방법은 음절로부터 음절의 길이에 영향이 적은 영역을 정하고 거기에서 음절의 특징을 추출하는 방법이다. 본 방법의 장점은 기존 신경망에서처럼 음절의 특징 파라메타를 추출하기 위해 프레임을 만들 때 윈도우의 길이를 음절의 길이에 따라 변화시키지 않아도 된다는 점이다. 이러한 음절의 특징을 추출하기 위해 한국어 음절의 특징을 이용하였다.

한국어 음절은 초성과 중성 그리고 종성으로 이루어져 있다는 특성이 있다. 이 같은 특성을 고찰해 보면 초성과 종성은 자음으로, 중성은 모음으로 이루어져 있고 이 점을 이용하면 음절의 길이에 영향을 적게 받는 특징을 추출 할 수 있음을 알게 된다.

소리를 발음할 때 있어서 자음은 모음과 결합할 때만 발음이 된다. 즉 음절을 다음과 같이 초성+중성, 중성+종성인 두 영역으로 나누었다고 할 때 초성+중성 영역에서 초성인 자음은 중성인 모음과 결합할 때만 발음할 수 있고 자음 자체로는 발음되는 구간이 한정되어 있다고 볼 수 있다. 따라서 초성인 자음은 음절 길이의 변화에 대해 일정하며 중성인 모음은 음절 길이의 변화에 가장 큰 영향을 끼치게 된다. 또한 모음의 음성 파형은 그 패턴이 규칙적이며 어는 한 부분이나 전체나 주파수 특성이 비슷하게 나타난다. 이러한 사실로 볼 때, 음절 시작의 일정 구간과 음절의 끝 일정 구간만 가지고도 그 음절이 무엇인지 알 수 있게 된다.

본 논문에서는 다음과 같은 사실로부터 한 음절에서 일정한 크기의 특징 벡터를 추출할 수 있었다.

■ 음절의 길이는 모음에 의해서 가장 큰 영향을 받는다.

■ 음절의 시작 구간 일정 영역과 끝 구간 일정 영역으로 그 음절이 무엇인지를 알 수 있다.

이러한 사실을 검증하기 위해서 다음과 같은 실험을 해 보았다.

하나의 단음절 숫자음을 길이가 틀리게 각각 발음한 뒤 음절의 시작 구간 일정 영역과 끝 구간 일정 영역을 취하여 이 두 영역의 음성신호를 합성한 뒤 차이가 있는가를 알아보았다.

그림 1은 이 실험에 쓰인 음성 파형 중 발음 길이가 각각 틀린 숫자음 '7'의 음성 파형 3개를 보인 것이다.

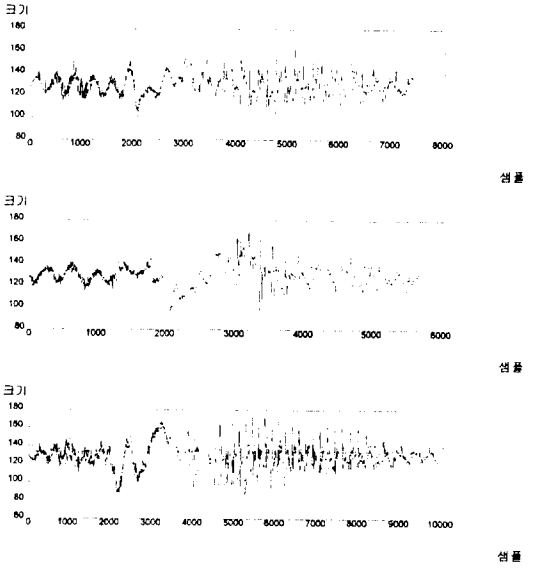


그림 1. 숫자음 '7'의 음성 파형
(a) 7449 샘플 (b) 5655 샘플 (c) 9870 샘플
Fig. 1. The waveform of the speech signal of "seven".
(a) 7449 samples (b) 5655 samples (c) 9870 samples

이와 같이 각각 길이가 다를 음성 신호를 신경회로망을 이용하여 인식을 하는 경우에는 신경회로망의 입력 벡터의 수가 일정해야 함으로 프레임 수를 일정하게 만들어 주기 위해서 각 음성신호마다 윈도우의 크기를 바꾸어 주어야 하는 단점이 생기게 된다.

본 논문에서는 이러한 단점을 해결하기 위해 일정 크기의 윈도우를 씌워 프레임을 구성하는 방법을 제안해 사용했다. 각 음성마다 각각 프레임의 개수가 다르게 되지만 한국어 음절의 특징을 이용하여 초성과 중성 영역의 일정 개수의 프레임에서 특징을 추출하여 MLP1으로, 중성과 종성 영역의 일정 개수의 프레임에서 특징을 추출하여 MLP2로 학습 및 인식을 시켰다.

그림 2에는 이러한 가정을 증명하기 위해서 그림 1에서 보인 각각 길이가 틀린 숫자음으로부터 초성과 중성 영역, 중성과 종성 영역에서 각각 2800 샘플의 신호를 얻어 합성 시킨 숫자음을 보였다. 세 개의 합성된 음성은 거의 비슷한 파형을 보이고 있고, 직접 숫자음을 들어 본 결과 모두 '7'로 발음이 됨을 확인 할 수 있었다.

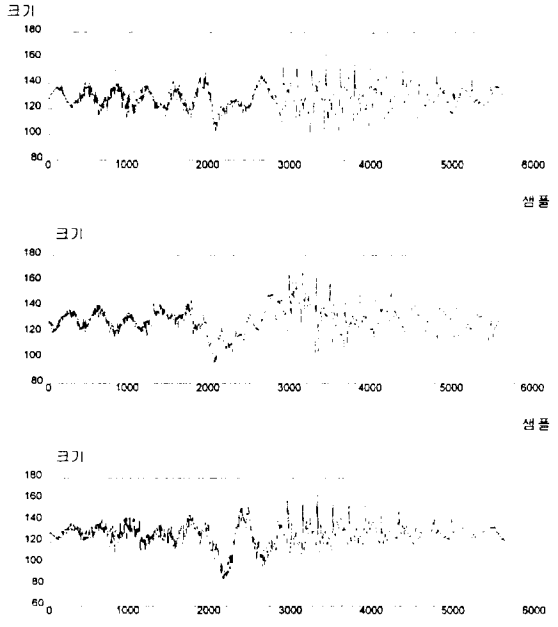


그림 2. 숫자음 '7'의 합성된 음성 파형 (5600 샘플)
Fig. 2. The waveform of the synthetic speech signal of 3 "seven" (5600 samples).

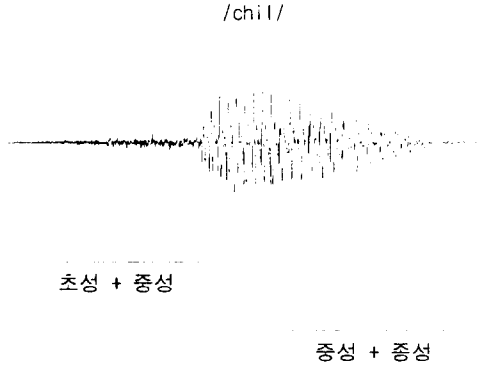


그림 3. 특징벡터 추출 영역
Fig. 3. The parts of the feature extraction.

본 논문에서는 이러한 한국어 음절의 특징을 이용하여 얻은 음절의 정적인 영역에서 특징 벡터를 추출할 수 있었고 이것은 음절의 길이에 영향이 작은 특징 벡터이기 때문에 음성의 프레임의 개수에 관계 없이 MLP로 학습 및 인식을 할 수 있었다. 이러한 특징벡터 추출 구간을 그림 4.3에 보인다.

2. Karhunen-Loeve 변환

1) KL변환의 특성

본 논문에서 MLP의 전처리로 사용한 KL변환은 요소(factor)분석의 특별한 방법으로, 서로 상관되어 있는 랜덤 변수들을 분석하기 위해 사용되는 수학적인 방법이다.

KL변환에 의해 측정 패턴으로부터 비상관적인 특징(feature)을 추출하는 선형 변환이 가능하게 된다. 그 이유는 이 변환에 의해서 각 그룹(class)간에 분산-상호분산 행렬(variance-covariance matrix)이 대각화(diagonal)되기 때문이다. 이와 같은 이유로 KL변환은 가장 최적의 변환으로 알려져 있다.

이러한 KL변환은 다음과 같은 특징이 있다.

■ 각각의 신호들 간의 상관성을 제거한다.

■ 신호의 차원 수를 감소시킨다.

MLP의 특성상 학습할 데이터의 크기가 커지면 입력 노드의 수가 증가하여 학습 시 시간이 너무 많이 걸리게 되고, 그렇게 되면 최적의 MLP 구조를 결정하는 데에도 더욱 많은 시행착오를 겪게 된다. 또한 입력 노드의 수가 증가할수록 가중치 벡터의 크기도 증가해서 메모리 사용량이 많아질 뿐만 아니라, 학습 시 국부 최소점으로 수렴할 확률도 그만큼 커지게 된다. 이러한 점으로 미루어 볼 때 KL변환을 사용하여 특징 벡터의 크기를 줄여 준다는 것은 대단한 효용성이 있는 것이라 할 수 있다. MLP는 오류 역전파(back propagation) 방법을 사용하여 학습을 하게 되는데 이때 학습되어야 할 패턴들의 상관성이 크다면 학습 시 속도가 느려지게 된다. 그 이유는 학습하기 위해 가중치 벡터를 변화시킬 때 오차가 최소화되는 경사로 가중치 벡터를 변화시키기 때문인데, 가중치 벡터의 한 요소의 변화는 출력 벡터인 목표하는 패턴의 모든 요소에 크고 작은 영향을 끼치게 된다. 따라서 학습할 패턴 간에 상관성이 클수록 다른 패턴의 목표 패턴에 상호 간섭하는 경향이 크게 되며 상관성이 적게 되면 그런 경향이 작아지게 된다. KL변환은 패턴들 간에 상관성을 제거시켜 주기 때문에 MLP의 학습 시 더욱 빠른 수렴을 보일 수 있다고 할 수 있다.

2) KL 변환 행렬 구하기

KL 변환은 다음과 같은 식에 의해 변환 행렬을 구할 수 있다.

각각의 숫자음의 LPC 계수를 이라고 할 때 패턴 집합의 벡터를 다음과 같이 표시한다.

$$x = [x_1 x_2 \cdots x_n]^T \quad (1)$$

그리고 이것의 평균을 구하면

$$m_x = E\{x\} \quad (2)$$

로 주어질 수 있고, 위의 두 식에서 의 상호분산 (covariance) 를 구하면

$$C_x = E\{(x-m_x)(x-m_x)^T\} \quad (3)$$

과 같은 식이 된다.

의 고유 벡터와 그에 상응하는 고유치를 ei , λ_i 라고 하고, $\lambda_i \geq \lambda_{i+1}$ $j=1,2,\ldots,n-1$ 이 되도록 고유치를 배열하고

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0 \quad (4)$$

각각의 고유치에 해당하는 고유 벡터를 변환 행렬A의 행이 되도록 순서대로 나열한다. 즉 변환 행렬 A의 첫번 째 행은 가장 큰 고유치에 해당하는 고유 벡터가, 그리고 마지막 행에는 가장 작은 고유치에 해당하는 고유 벡터 가 오도록 한다.

$$A = [e_1 \; e_2 \; \ldots \; e_n]^T \quad (5)$$

라는 벡터에서 라는 벡터로 변환하는 행렬을 라고 하면

$$y = A(x-m_x) \quad (6)$$

과 같은 식이 된다.

행렬의 위쪽의 행에 의해 변환된 값들은 변환될 패턴들의 중요한 성분에 해당 된다. 그리고 이때 변환된 값 는 상호분산이 0이 된다. 그리고 에너지가 작은 특징에 해당하는 λ_i가 작은 성분을 무시함으로써 의 행의 크기를 줄일 수 있다. 즉, 패턴의 주요 요소들은 λ_i가 큰 것에 해당하므로 크기가 큰 몇몇 λ_i에 해당하는 것 외에 것을 소거함으로써 행렬의 크기를 줄일 수 있고 따라서 그 행렬에 의해 변환된 입력 패턴의 크기도 줄일 수 있다.

3) LPC계수의 KL 변환

11 KHz로 샘플링 된 숫자음을 400샘플 크기의 해밍 윈도우를 씌우고 200샘플씩 이동하며 차례대로 프레임을 구한 뒤에 각각의 프레임에서 16차 LPC계수를 추출하게 되며 숫자음의 시작 영역 7레임과 끝 영역 7프레임의 LPC계수를 각각 초성과 중성 영역과 중성과 종성 영역의 특징 벡터로 하였다. 이 특징 벡터들 중 학습할 0부터 9까지의 숫자음의 특징 벡터로 부터 KL변환 행렬을 구하고 변환 행렬에 곱해진 특징 벡터는 차원이 감소하게 되어 크기가 줄어든다.

그림 4과 5에는 숫자음 '0'의 시작 부분의 LPC계수 7프레임과 그것에 KL변환 행렬을 곱한 결과를 보인 것이다. KL변환을 만들 때 사용한 데이타는 학습에 참가한 150개의 숫자음이였고, KL변환 행렬의 크기는 112x112로서 차원을 감소시키지 않은 행렬을 사용한 결과를 그림 4.4에 나타내었다. 각 그림은 세 명의 화자로부터 얻은 15개의 숫자음 '0'에 관한 LPC계수 및 그것의 KL변환한 특징으로서 각각의 패턴을 겹쳐서 나타내었다.

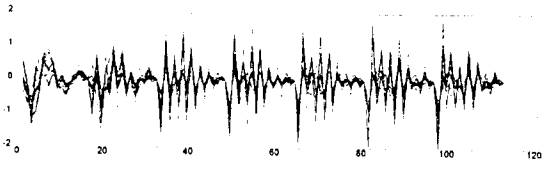


그림 4. 숫자음 0의 LPC 계수
Fig. 4. LPC coefficients of speech '0'.

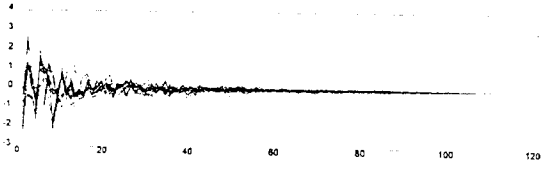


그림 5. LPC 계수를 변환한 결과
Fig. 5. Transfored LPC coefficients.

KL변환은 복잡한 계산 과정이 필요하지만 인식할 때에는 그러한 계산 과정이 필요한 것이 아니라 단지 변환 행렬을 곱해 줌으로써 KL변환 한 효과를 갖게 된다. 그것은 학습한 데이타와 인식할 데이타의 통계적 특성이 서로 비슷하기 때문이다. 또한 KL변환에 의해 상관성이 없어진다 하더라도 같은 숫자음은 비슷한 패턴으로 결과가 나오게 된다. 그림 6에는 학습에 참가하지 않은 숫자음 '0'의 LPC계수를 보이고 그림 7에는

그 LPC계수를 KL변환 행렬에 곱한 결과를 나타내었다.

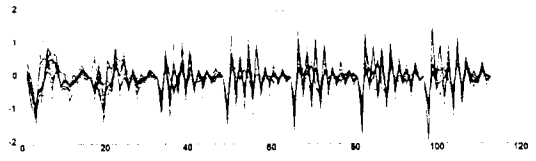


그림 6. 숫자음 0의 LPC 계수 (인식에 사용된 것)
Fig. 6. LPC coefficients of speech '0' (used in the recognition).

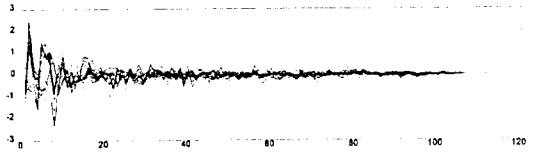


그림 7. LPC 계수를 변환한 결과 (인식에 사용된것)
Fig. 7. Transformed LPC coefficients (used in the recognition).

LPC계수에 KL변환 행렬을 곱한 결과를 보면 처음 8부분의 에너지가 크고 나중으로 갈수록 에너지가 작아지는 것을 볼 수 있다. 그것은 가장 큰 특징들을 앞으로 배열했기 때문인데 이런 이유로 뒷 부분의 에너지가 작은 부분들을 무시함으로써 에너지가 큰 패턴들, 즉 중요한 특징을 대표하는 패턴들만을 선택하기 때문에 특징의 크기를 줄일 수 있다. 그림 8과 그림 9에는 각각 학습 및 인식에 사용한 LPC계수를 KL변환시킬 때, 변환 행렬에서 10개의 행만을 사용하여 패턴의 크기를 줄인 것을 나타내었다. 이것은 그림 4와 6에서 KL변환 된 결과에서 10번째까지의 샘플만을 나타낸 것과 동일한 것이다.

그림 8과 9에서 학습한 데이타의 패턴과 인식에 사용될 데이타의 패턴이 유사함을 알 수 있다.

KL 변환은 크고 복잡한 패턴들에서 각각의 중요한 특징들만을 추출하여 크기를 줄여 주어 간단하면서도 특징 있는 패턴으로 만들어 주지만 그러한 과정에서 변환하기 전에 가지고 있는 물리적 특성 및 시간적 상관성을 잃어버리게 된다. 따라서 KL변환은 TDNN과 같은 시간적 상관성을 반영하는 신경회로망에는 사용하기 힘들 것이고 본 연구에서 사용한 MLP는 정적인 패턴을 인식하는데 편리하기 때문에 KL변환을 전처리로 사용하기에 적당했다.

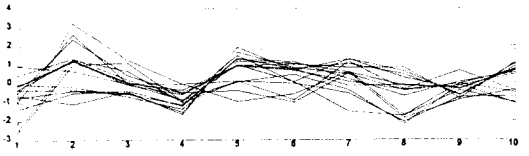


그림 8. KL변환 결과의 중요한 특징
Fig. 8. Compressed feature using KLT.

그림 9. KL변환 결과의 중요한 특징 (인식에 사용된 것)
Fig. 9. Compressed feature using KLT. (used in the recognition).

하지만 단음절인 숫자음 인식에 적용한다 하더라도 음절의 길이의 변화로 인한 시간적 왜곡 및 연속 음에서의 조음 현상 등의 동적인 특성을 갖는 음성을 이러한 KL변환과 MLP를 사용하여 인식하기 위해서 인식시에 동적인 특성을 고려한 인식 기법이 요구 되었다.

Ⅲ. 숫자음 인식 기법

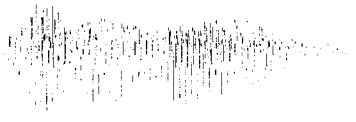
학습 시에 숫자음의 앞 부분과 뒷 부분으로 각각 나누어 학습했기 때문에 예를들면, 3과 4 또는 1과 2, 0과 6등은 앞 부분이 서로 비슷하지만 뒷 부분이 틀리고 1과 7 또는 5와 9는 앞 부분이 틀리고 뒷 부분이 서로 비슷하다는 정보를 참조해 인식을 수행하게 된다 [표 1].

표 1. 숫자음 인식의 최종 판별을 위한 표
Table 1. Decision table for final recogmition.

MLP1	MLP2	최종 인식
0, 6	0	0
0, 6	6	6
1	1, 7, 8	1
1, 2	1	1
3, 4	3, 0	3
3, 4, 8	4	4
3, 4, 8	1, 7, 8	8
5	5, 9	5
9	5, 9	9

1. 연속 숫자음 인식

연속 숫자음 인식시에는 숫자음과 숫자음이 결합하는 데서 나타나는 조음 현상 및 음절 사이의 구분이 힘들다는 등의 많은 어려움이 있다. 만일 각각의 음절이 명확히 구분되어질 수 있다면 고립 숫자음을 인식할 때와 마찬가지로 인식하면 된다. 하지만 음절의 구분이 모호한 연속 숫자음의 경우에는 고립 숫자음의 경우와는 달리 MLP1과 MLP2에서의 예측 및 결과를 통해서 인식하게 된다.



MLP1 *1111111111111111*5556
MLP2 44444*1117*440747*2222222*
Energy 2334433443333333333333

그림 10. 연속 숫자음 인식 예
Fig. 10. Example of continuous numeric speech recognition.

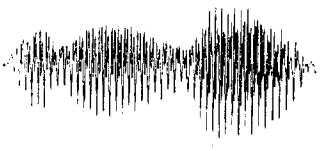
그림 10에는 음절의 구분이 모호한 숫자음 '1 2(일이)'의 연속음의 파형을 보이고 인식 과정을 나타내었다. 에너지의 숫자열은 각 프레임의 에너지를 대수화 시킨 것이다. 이 숫자음은 음절 구간의 분리가 쉽지 않기 때문에 기존의 고립어로 학습시킨 인식 시스템에서는 인식하기가 힘들다. 하지만 본 연구에서는 고립어로 학습을 시킨 MLP로 연속음 인식에 적용할 수 있다는 것을 보이고 있다.

인식과정

1. 시작점을 검출하고 특징점을 추출한다.
2. 프레임의 개수가 6이 넘어가면 6프레임씩 차례로 인식한다

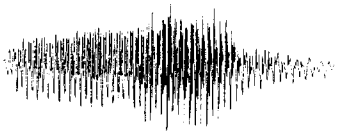
(고립어 인식에서는 7프레임을 사용하였으나 연속음에서는 발음 속도가 다소 빨라져서 6프레임 단위로 인식을 하게 되었다).

3. 음절의 시작 부분은 MLP1으로인식하면서 예상되는 숫자음을 결정한다. 예상되는 숫자음의 결정 방법은 1 다음과 같다.
   a. 숫자음의 발음이 시작됨을 에너지의 크기로 판단하고 인식을 시작한다.
   b. 인식을 차례로 하면서 비슷하게 연속으로 결과가 나오는 경우 그것을 예상되는 숫자음으로 결정한다.
   c. 뒤 따라 인식되는 MLP2의 결과 및 음성의 에너지 크기를 고려해서 MLP1과 조합 가능한 숫자음 인지를 결정한다.
4. MLP1과 MLP2와의 상호 관련성이 있다면 예상되는 숫자음을 최종 인식된 숫자음으로 판단한고 그렇지 않은 경우는 인식을 실패한 것으로 판단한다.
5. MLP1에서 새로운 숫자음 후보를 찾는다. 3-4를 반복한다.



MLP1 *111111*12888*66000*
MLP2 4*77777*4442711*666*
Energy 33333333333333333

그림 11. 연속음 인식 예
Fig. 11. Example of continuous numeric speech recognition.



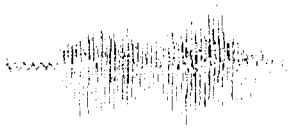
MLP1 *22222222222*6*000000*998
MLP2 *2222222222*444444*00000*
Energy 22333333333334444433

그림 12. 연속음 인식 예
Fig. 12. Example of continuous numeric speech recognition.

Ⅳ. 실험 및 결과

1. 연속 숫자음 인식

연속 숫자음의 학습 및 인식 실험의 데이타는 단일 화자가 10번씩 발음한 100개의 고립 숫자음을 학습용으로 사용하고 인식 실험은 동일 화자가 발음한 200개의 두자리 연속 숫자음을 대상으로 하였다. 연속 숫자음은 자연스럽게 이어서 발음했으며 다소 잡음이 심한 환경에서 8 비트의 양자화 레벨로 데이타를 얻었다.



MLP1 99999955556666600030009999
MLP2 99977447744411140000000000
Energy 22222233333333333344444333

그림 13. 연속음 인식 예
Fig. 13. Example of continuous numeric speech recognition.

인식 결과는 표 2부터 표 4에 나타내었다. 대부분의 오차는 숫자음 '4'와 '3'에서 발생하였으며 특히 '4'로 시작하는 연속 숫자음의 경우 MLP2에서 인식을 잘 못하는 경우가 많았는데 그것은 조음 효과로 인하여 '4'의 초성과 중성 부분의 발음이 학습 데이타였던 고립어와 달랐기 때문이라고 추측 된다.

표 2. 첫번째 숫자음의 인식률
Table 2. The recognition rate of the first.

첫번째 숫자음	오인식 갯수	인식률(%)
0	1	95
1	2	90
2	0	100
3	2	90
4	4	80
5	2	90
6	1	95
7	4	80
8	5	75
9	1	95
TOTAL	22	89

표 3. 두번째 숫자음의 인식률
Table 3. The recognition rate of the second syllable.

두번째 숫자음	오인식 갯수	인식률 (%)
0	1	95
1	1	95
2	0	100
3	3	90
4	6	70
5	2	90
6	1	95
7	4	80
8	1	95
9	2	90
TOTAL	20	90

표 4. 두자리 연속음의 인식률
Table 4. The recognition rate of 2-digits continuous words.

연속 숫자음	TOTAL
두 숫자 모두 맞은 갯수	162
인식률 (%)	81

V. 결 론

기존의 신경회로망인 MLP는 단순한 구조 및 빠른 인식 속도를 갖는 정적인 패턴 인식에 유용한 패턴 분류기이다. 하지만 음성신호와 같은 동적인 패턴을 인식하는 데에는 여러 문제점이 있으며 음절 단위의 인식을 시도할 때에는 효과적으로 적용할 수 있으나 시간적인 왜곡 등을 처리하기 위해서 인식할 음절의 전체 프레임 개수를 일정하게 만들어 주어야 하여 이를 위해서는 프레임을 만드는 윈도우의 길이를 가변 시켜 주어야 하는 단점이 있다. 또한 음절의 구분이 명확하지 않은 연속 음에는 적용하기 힘들다.

또한 기존의 MLP에서는 음성신호의 막대한 크기의 데이타를 처리하기 위해 입력 노드의 수를 크게 할 수 밖에 없었다.

이로써 신경망이 커지며 계산량이 많고 학습 시간이 긴 단점이 있다.

본 논문에서는 한국어 음절의 특성을 이용하여 음절의 앞 부분과 뒷 부분의 일정 영역에서 특징을 추출하고 각각의 MLP로 학습 및 인식을 함으로써 일정한 윈도우로 프레임을 만들면서도 음절의 길이 변화에 잘 적응하는 모델을 제시함으로써 MLP만으로도 시간적 상관성을 반영하는 TDNN과 같은 신경망을 사용한 효과를 얻을 수 있었으며 기존의 신경망으로는 구현하기 힘든 음절의 구분이 모호한 연속음의 음절 단위의 인식에도 적용될 수 있음을 보였다.

또한 본 논문에서는 학습할 패턴들을 압축을 시켜주는 전처리 과정을 도입하였고 이 전처리 과정으로서 Karhunen-Loeve 변환을 사용하는 방식을 제안하였다. 이 전처리 과정을 통해서 빠른 학습 속도를 얻었고 전체적인 사용 메모리 크기를 감소시킬 수 있었다.

본 연구는 한국어 숫자음 만을 대상으로 하였으나 앞으로 한국어의 모든 음절을 대상으로 한 연구가 필요하다.

참 고 문 헌

[1] Gonzalez, Rafael C., *Digital image processing*, Addison Wesley, 1992.

[2] 정인길, "시간 지연 쌍전파 신경회로망을 이용한 격리 단어의 인식," 인하대학원 전자공학과 석사 학위 논문, 1995년 2월

[3] 이영호, 정홍, "음절을 기반으로한 한국어 음성인식," 전자공학회논문지, 제 31권, B편, 제 1호, 1994. 1., pp. 11-22

[4] 이종식, "시공간 신경회로망을 이용한 한국어 숫자음 인식," 인하대학원 전자공학과 석사학위 논문, 1995년 8월

저 자 소 개

朴 廷 緖(正會員)
1971년 10월 9일생. 1994년 2월 인하대학교 전자공학과 학사. 1996년 2월 인하대학교 대학원 전자공학과 석사. 1996년 2월 ~ 현재 픽셀 시스템 연구원. 주관심 분야는 음성인식, 신경회로망

權 五 常(正會員)
1967年 10月 17일생. 1990年 2月 인하대학교 전자공학과 졸업(공학사). 1992年 2月 인하대학교 대학원 전자공학과 졸업(공학석사). 1992年 1月 ~ 1996년 2월 대우중공업 중앙연구소 전자기술실 주임연구원. 1995年 3月 ~ 현재 인하대학교 대학원 전자공학과 박사과정. 주관심 분야는 생체신호처리, 재활 로보틱스

洪 勝 弘(正會員) 第 32卷 B編 第 7號 參照

權 章 禹(正會員) 第 32卷 B編 第 7號 參照

李 應 赫(正會員) 第 32卷 B編 第 10號 參照