論文95-32B-12-13

신경회로망과 Markov 모델을 이용한 한국어 속담 인식에 관한 연구

(A study on the Recognition of Korean Proverb Using Neural Network and Markov Model)

洪 基 源 *, 金 善 一 *, 李 幸 世 *

(Ki Won Hong, Seon Il Kim, and Haing Sei Lee)

요 약

본 논문은 신경회로망과 Markov 모델을 이용한 한국어 속담 인식에 관한 논문이다. 음성의 특징추출은 5ms 마다 PLP(Perceptual Linear Prediction) 분석을 통하여 PLP-Cepstrum 계수를 추출하고 여기에 영교차율, 단구간 에너지를 포함하여 음성의 한 프레임으로 구성하였다. 음소인식을 위한 신경회로망의 입력으로 사용되는 입력벡터는 음소간의 관계 학습을 위하여 300ms의 데이타를 사용하였다. 단어 및 속담 인식을 위해서는 음소열과 단어열을 Markov 모델에 적용하여 인식을 하였다. 인식 실험 결과 음소의 경우는 81.2%의 인식율과 단어의 경우는 94.0%의 인식율을 얻었다.

Abstract

This paper is a study on the recognition of Korean proverb using neural network and Markov model. The neural network uses, at the stage of training neurons, features such as the rate of zero crossing, short-term energy and PLP-Cepstrum, covering a time of 300ms long. Markov models were generated by the recognized phoneme strings. The recognition of words and proverbs using Markov models have been carried out. Experimental results show that phoneme and word recognition rates are 81.2%, 94.0% respectively for Korean proverb recognition experiments.

Ⅰ. 서 론

인간이 의사를 전달하는 가장 기본적인 도구는 문자와 음성이다. 의사전달 수단중에서 기록으로 남기기 위한 도구로서 문자가 존재하지만 가장 쉽고 자주 사용되는 것이 음성을 통한 정보의 교환이다. 최근에는 인간 사이의 정보 교환외에 인간과 기계 사이의 정보 교환도 중요한 상황이 되었다. 또한, 이를 위해 많은 연구가 이루어져 왔다. 인간과 기계사이의 정보 교환에는 많은 도구를 사용할 수 있지만, 그 중에서도 가장 자연스러운 것이 음성이다. 기계에 음성을 인식시키기 위해 여러가지 방법이 개발되어 왔고, 음성을 가장 잘 표현해줄 수 있는 특징을 찾기위한 노력이 이루어져 왔다. 그러나 기계가 인간과 같은 능력으로 음성을 인식하기에는 아직은 초보적인 단계에 불과하다. 따라서 지금까지는 대상 어휘를 제한하여 고립단어나 숫자[9,10] 등의 제한 영역에서 응용하고자 하는 노력이 이루어져 왔다.

음성 인식 방법에는 많은 방법이 사용되고 있다. 가장 널리 사용되는 방법에는 DTW(Dynamic Time

* 正會員, 亞洲大學校 電子工學科

(Department of Electronics Engineering, Ajou University)

※ 본 논문은 1995년도 아주대학교 교내연구비 지원에 의해 연구된 것입니다.

接受日字:1995年10月6日, 수정완료일:1995年11月22日

Warping), HMM(Hidden Markov Model), ANN (Artificial Neural Network) 등이 사용된다. 최근에는 ANN과 HMM을 결합하여 음성을 인식하는 방법이 사용되고 있다.

본 논문에서는 신경회로망과 HMM에서 사용되는 Markov 모델을 이용하여 연속음성에서의 음소, 단어, 속담을 인식하는 시스템을 구성하였다. 그리고 음성신호의 특징을 추출하기 위하여 인간의 청각 신경을 모방한 인지선형예측법(PLP : Perceptual Linear Prediction)[6,7,8]을 이용하여 5ms 마다 음성의 특징을 추출하였고, 연속음성에 나타나는 음소간의 관계 학습을 위하여 300ms의 입력벡터를 구성하여 신경회로망의 입력으로 사용하였다. 특징추출에 의한 음성의 특징값은 오차 역전달 알고리즘을 이용한 MLP 신경회로망을 이용하여 연속음성에 나타나는 음소를 인식하고, 신경회로망을 통해서 인식된 음소열은 Markov 모델에 적용하여 각 단어에 대한 통계값을 구한뒤 연속음성에 나타나는 단어를 인식하였다. 속담 인식은 단어 인식에서 인식된 단어열을 바탕으로 Markov에 적용하여 각 속담에 대한 통계값을 구한 뒤 속담을 인식하였다.

표 1. 실험 데이타(한국어 속담)
Table 1. Experimental data(Korean Proverb)

D01	가자니 태산이요 돌아서자니 숭산이라
D02	고래 싸움에 새우등 터진다
D03	구르는 돌에는 이끼가 끼지 않는다
D04	낙숫물이 댓돌을 뚫는다
D05	눈뜨고 도둑 맞는다
D06	달면 삼키고 쓰면 뱉는다
D07	미련은 먼저나고 슬기는 나중난다
D08	새벽달 보자고 초저녁부터 기다린다
D09	여우를 피하면 호랑이를 만난다
D10	엎구리 찔러 절받기
D11	지렁이도 밟으면 꿈틀거린다
D12	지성이면 감천이다
D13	집신장수 헌신만 신는다
D14	천리길도 한걸음부터
D15	칼로베고 소금친다
D16	큰방죽도 개미구멍으로 무너진다
D17	태산을 넘으면 평지를 본다
D18	핑계없는 무덤없다
D19	호미로 막을걸 가래로 막는다
D20	흐르는 물도 떠주면 공덕이다

Ⅱ. 음성의 특징추출

실험에서 사용되는 음성 데이타는 음성보드 DT2801을 이용하여 10kHz, 12bits로 양자화된 한국어 속담을 사용하였다. 표 1은 실험에 사용된 20개의 한국어 속담 데이타이다. 20개의 한국어 속담에는 309개의 음소와 67개의 단어로 구성되어 있다. 실험에 사용된 음소는 속담에 나타나는 음소만을 가지고 사용하였다. 음소의 구성은 초성 자음 'ㄱ', 'ㄲ', 'ㄷ', 'ㄸ', 'ㅂ', 'ㅅ', 'ㅆ', 'ㅈ', 'ㅉ', 'ㅊ', 'ㅋ', 'ㅌ', 'ㅍ', 'ㅎ' 과 종성 자음 'ㄱ', 'ㄷ', 'ㅇ' 그리고 'ㄴ', 'ㄹ', 'ㅁ' 은 초성이나 종성에서 비슷한 음가를 가지므로 구별하지 않고 하나로 사용하였다. 그리고 모음은 'ㅏ', 'ㅓ', 'ㅕ', 'ㅗ', 'ㅛ', 'ㅜ', 'ㅠ', 'ㅡ', 'ㅣ', 'ㅐ,ㅒ,ㅔ' 로 이루어진 모음을 사용하였고 묵음을 포함하여 전체의 음소를 구성하였다. 신경회로망 학습을 위한 음소의 결정은 연속음성에서 나타나는 각 음소에 대한 음가를 기본으로하여 음소를 결정하였다.

음성의 특징추출은 5ms 마다 PLP 분석을 통하여 추출된 7차의 PLP-Cepstrum과 단구간 에너지, 영교차율을 사용하여 해당 음성을 표현하는 하나의 음성 프레임으로 구성하였다. 음성의 특징추출 과정은 그림 1과 같다.

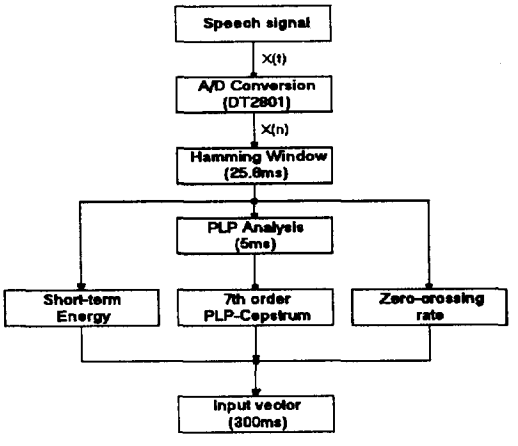

그림 1. 음성의 특징추출 과정
Fig. 1. The procedure of feature extraction.

PLP 분석법[6,7,8]은 실제적인 공학적 근사에 의해 귀의 성질을 모방하고, 귀가 느끼는 스펙트럼은 자기회귀 전극점 모델로 근사된다. PLP 분석법은 그림 2에 나타낸 과정으로 처리된다.

PLP 분석을 통한 PLP 모델은 LP(Linear Prediction) 모델에 비해 더 낮은 차수의 전극점 모델의 스펙트럼에 의해 근사된다. 귀가 느끼는 스펙트럼은 0-5kHz의 주파수 범위에서 16개의 대역들로 나뉘어

서 합하여지며, 중간대역과 상위대역을 보강하기 위하여 equal-loudness pre-emphasis를 거치게 된다. 또한 음성 스펙트럼의 전력 변화율을 감소시키기 위하여 3 제곱근 처리를 함으로써 intensity-loudness 3 제곱근 크기 압축을 실시한다. 이러한 처리를 거친 16개의 스펙트럼 성분에 푸리에 역변환 과정을 적용시켜 자기 상관 계수를 얻는다. 전극점 모델은 얻어진 자기 상관 계수로부터 원하는 차수로 계산되며, 이로부터 다시 켑스트럼 계수를 계산할 수 있다.

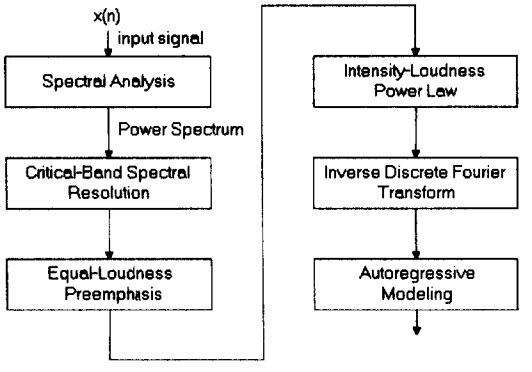

그림 2. PLP 분석법의 블럭도
Fig. 2. Block diagram of PLP speech analysis method.

PLP 분석을 위한 연산 복잡도는 LP분석에 비해 대단히 크다. 연산측면에서 가장 복잡한 부분이 FFT 스펙트럼 계산과, 이어지는 임계대역 스펙트럼 적분과 3 제곱근 압축등이다. AR 모델을 위한 연산의 복잡도는 주파수 분해능이 낮아져서 서로 상쇄될 수 있다.

PLP 모델의 결과 스펙트럼은 일반 LP 모델의 스펙트럼에 비해 더 선형적이다. 또한 일반 LP 모델에 비해 더 낮은 차수의 모델링이 가능하다. 결과적으로 신경회로망의 입력의 감소와 데이타베이스의 역할을 하는 가중치들을 줄이는 역할을 하게되며, 이것은 처리속도의 효율화에 기여하게 된다.

Ⅲ. 제안하는 인식 시스템

제안하는 음성인식 시스템은 신경회로망과 Markov 모델을 이용하여 구성하였다. 인식 시스템의 구성은 그림 3에 나타낸것과 같다.

인식 시스템에서는 음소 인식을 위해 MLP 신경망을 사용하였고, 이를 통해 연속음성에 나타나는 음소를 인식하고, 인식된 음소열을 Markov 모델에 적용하여 각 단어 및 속담에 대한 통계값을 계산하여 단어 및 속담에 대한 인식을 수행하게 된다.

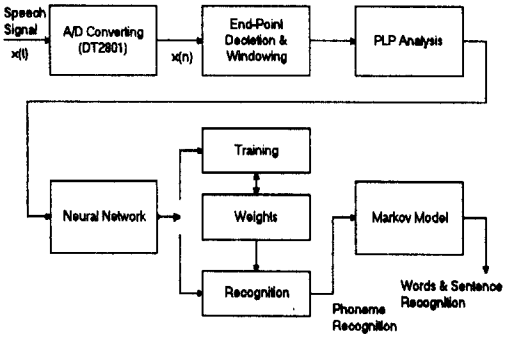

그림 3. 인식 시스템의 블럭도
Fig. 3. Block diagram of recognition system.

1. 입력 벡터의 구성

신경회로망의 학습을 위해서 먼저 5ms 마다 PLP 분석에 의한 특징값들은 각 음소에 대하여 사람의 손에 의해서 음소로 분할된 데이타를 구성하게 된다. 이런 음소 분할된 데이타를 이용하여 입력 벡터을 구성하게 되는데, 이는 음소간의 관계학습을 위해 구성하였다. 연속음에 나타나는 음소는 전후의 다른 음소와의 조음 현상에 의한 변화가 심하다. 현재의 음성 조각을 인식하기 위해서는 인접한 음소를 같이 참조하는 것이 타당하다. 보통은 이러한 변화를 학습하기 위하여 시간 지연신경망을 사용하지만 본 논문에서는 고정 신경망을 사용하고, 입력단의 구성을 달리하여 시간 변화를 학습하였다. 5ms의 음성 프레임을 인식하기 위하여 특징들의 입력 벡터를 구성하였다. 입력 벡터의 구성은 그림 4와 같다.

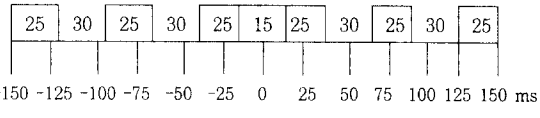

그림 4. 입력 벡터의 요소 구성
Fig. 4. The distribution of input vector.

네모칸 안의 숫자는 각 블럭의 시간폭을 나타낸다. 실선은 음성의 특징 값이 평균되어지는 구간의 길이를 나타내며, 점선은 사용되지 않는 구간을 나타낸다. 네모칸 안의 15는 현재의 프레임(5ms) 및 전 프레임(5ms), 후 프레임(5ms), 총 3개의 프레임으로 한 블럭을 형성한 것을 나타내며 세 프레임에서 구한 특징

들이 평균되어 한 특징 벡터로 나타내어진다. 25는 5개의 프레임(총25ms)으로 한 블럭을 형성한 것을 나타내며 다섯 프레임에서 구한 특징들이 평균되어 한 특징 벡터로 나타내어진다. 30은 사용되지 않는 구간인데 6개의 프레임 즉 30ms 길이이다.

2. 신경회로망과 Markov 모델의 구성

신경회로망은 일반적으로 분류기능이 우수하다. 본 논문에서는 이러한 점을 감안하여 음소 인식을 위하여 사용하였다. 사용된 신경회로망은 오차역 전달 알고리즘을 이용한 MLP 신경망을 사용하였다. 신경회로망의 입력은 300ms 에 추출된 입력 벡터로 63개의 입력으로 구성되었고, 중간층은 40, 출력층은 음소 출력인 30개로 구성되었다. 비선형함수로는 시그모이드 함수를 사용하였다. 신경회로망에 비해 Markov 모델은 시간적 특성을 고려한 것으로 HMM에 비하여 구현이 쉬운 장점이 있다. 또한 단일 통계 모델이기 때문에 신경회로망의 출력인 음소를 단어나 속담으로 구성하는데 장점이 있기 때문에 단어 및 속담을 인식하는 후처리 과정으로서 사용하였다. 본 논문에서는 단어 인식시에는 음소열을 이용하고, 속담 인식시에는 단어열을 이용하여 인식을 수행하였다.

Ⅳ. 실험 및 검토

한국어 속담에 대한 인식 실험은 SUN SPARC 10에서 실험되었다. 신경회로망의 학습과 Markov 모델을 이용한 각 단어나 속담에 대한 기준 모델의 구성을 위하여 20대 남성 2인이 각각 두번씩 발음한 음성을 사용하였다. 그리고 인식 실험을 위해서 사용된 속담 데이타는 화자 종속과 독립적인 인식을 위해 20대 남성 각각 2인이 세번씩 발음한 음성을 사용하였다. 그리고 인식 실험은 연속음성에서 나타나는 음소, 단어, 속담에 대한 인식으로 나누어서 실험을 하였다.

1. 음소 인식 실험

음소 인식 실험은 연속음성에 나타나는 음소를 인식하기 위한 실험으로 신경회로망을 이용하여 인식을 수행하였다. 4명의 화자로부터 인식을 수행한 결과 평균 81.2%의 음소 인식율을 얻었다. 화자 종속과 화자 독립의 인식 결과를 표 2와 표 3에 나타내었다.

표 2와 표 3을 통한 음소 인식 결과에서 화자종속의 인식율은 84.4%의 인식율을 보였고, 화자독립의 인식율은 77.9%의 인식율을 보였다. 또한 모음에 비하여 자음의 경우에 낮은 인식율을 보였다. 이러한 이유로는 자음의 경우 비음과 'ㅅ', 'ㅆ'를 제외하고는 5ms 마다 음성을 분석하면 짧은 구간 동안에만 음성 프레임이 분포하여 음소 학습시 모음에 비하여 적게 학습되는 단점으로 모음에 비해 인식율의 저하를 가져왔다. 이러한 점은 음성 분석을 더 짧은 구간으로 한다면 다소 향상이 될 것이다.

표 2. 음소인식 실험 결과(화자 종속)
Table 2. Experimental results of phoneme recognition(speaker-dependent)

음 소		화 자 A	화 자 B	인 식 율
초성 자음	ㄱ	36/63	40/63	60.3%
	ㄲ	9/24	9/24	37.5%
	ㄷ	59/87	62/87	69.5%
	ㄸ	7/21	5/21	28.6%
	ㅂ	7/33	5/33	18.2%
	ㅅ	39/42	38/42	91.7%
	ㅆ	12/12	10/12	91.7%
	ㅈ	44/51	44/51	86.3%
	ㅉ	2/3	1/3	50.0%
	ㅊ	9/12	10/12	79.2%
	ㅋ	3/9	4/9	38.9%
	ㅌ	9/18	13/18	61.1%
	ㅍ	8/9	7/9	83.3%
	ㅎ	9/15	8/15	56.7%
초,종성	ㄴ	165/183	164/183	89.9%
	ㄹ	104/123	106/123	85.4%
	ㅁ	78/90	72/90	83.3%
종성	ㄱ	11/15	6/15	56.7%
	ㅂ	4/9	4/9	44.4%
	ㅇ	38/39	37/39	96.2%
모음	ㅏ	151/159	157/159	96.9%
	ㅓ	68/75	66/75	89.3%
	ㅕ	32/39	35/39	85.9%
	ㅗ	73/75	73/75	97.3%
	ㅛ	2/3	2/3	66.7%
	ㅜ	47/63	46/63	73.8%
	ㅡ	84/90	74/90	87.8%
	ㅣ	117/120	118/120	97.9%
	ㅔ	34/39	35/39	88.5%
묵 음		194/198	197/198	98.7%
인 식 율		84.6%	84.2%	84.4%

2. 단어 인식 실험

단어 인식 실험은 신경회로망을 통해서 인식된 음소

열을 바탕으로 이를 67개의 단어에 대한 Markov 모델에 적용하여 단어 인식을 하였다. 4명의 화자로부터의 인식 결과 화자 종속의 단어 인식에서는 96.8%의 인식율과 화자 독립 단어 인식에서는 91.3%의 인식율을 보였고, 평균적으로 94.0%의 단어 인식율을 얻었다. 단어 인식 결과는 표 4와 같다.

표 3. 음소인식 실험 결과(화자 독립)
Table 3. Experimental results of phoneme recognition(speaker-independent)

음 소		화 자 C	화 자 D	인 식 율
초성자음	ㄱ	41/63	38/63	62.7%
	ㄲ	6/24	4/24	20.8%
	ㄷ	28/87	31/87	33.9%
	ㄸ	3/21	4/21	16.7%
	ㅂ	7/33	4/33	16.7%
	ㅅ	30/42	34/42	76.2%
	ㅆ	6/12	8/12	58.3%
	ㅈ	41/51	44/51	83.3%
	ㅉ	1/3	1/3	33.3%
	ㅊ	6/12	8/12	58.3%
	ㅋ	2/9	2/9	22.2%
	ㅌ	7/18	8/18	41.7%
	ㅍ	6/9	7/9	72.2%
	ㅎ	11/15	7/15	60.0%
초,종성	ㄴ	132/183	136/183	73.2%
	ㄹ	106/123	100/123	83.7%
	ㅁ	76/90	69/90	80.6%
종성	ㄱ	7/15	6/15	43.3%
	ㅂ	3/9	2/9	27.8%
	ㅇ	36/39	28/39	82.1%
모음	ㅏ	154/159	151/159	95.9%
	ㅓ	66/75	60/75	84.0%
	ㅕ	26/39	27/39	67.9%
	ㅗ	64/75	66/75	86.7%
	ㅛ	2/3	2/3	66.7%
	ㅜ	40/63	41/63	64.3%
	ㅡ	72/90	62/90	74.4%
	ㅣ	111/120	117/120	95.0%
	ㅐ	33/39	34/39	85.9%
묵 음		193/198	193/198	97.5%
인 식 율		78.5%	77.3%	77.9%

인식 결과를 보면 속담에 나타나는 단어 중에서 "가래로", "가자니", "감천이다", "개미구멍으로", "공덕이다", "꿈틀거린다", "끼지", "낙숫물이", "눈뜨고", "댓돌을", "막을걸", "먼저나고", "무너진다", "무덤없다", "물도", "미련은", "뱉는다", "보자고", "삼키고", "새벽달", "새우등", "소금친다", "송산이라", "신는다", "싸움에", "쓰면", "엿구리", "절받기", "지렁이도", "지성이면", "짚신장수", "찔러", "천리길도", "초저녁부터", "큰방죽도", "태산이요", "평지를", "핑계없는", "한걸음부터", "헌신만", "호랑이를"은 화자 종속이나 화자 독립의 인식에서 100%의 단어 인식율을 보였다. 나머지 단어에서는 비슷한 음소의 구성으로 인한 약간의 오인식된 단어가 발생하였다.

표 4. 단어 인식 결과
Table 4. Experimental results of word recognition.

화 자		인식 갯수	인 식 율
화자종속	화 자 A	196/201	97.5%
	화 자 B	193/201	96.0%
화자독립	화 자 C	186/201	92.5%
	화 자 D	181/201	90.0%

단어 인식은 Markov 모델을 통해서 구성된 기준 단어를 이용하므로 음소 인식시 일부가 잘못 인식될 경우에도 기준 단어와 거리를 비교하여 최소 거리를 나타내는 단어로 인식한다. 따라서 음소에 비해 인식율이 향상을 가져왔다.

3. 속담 인식 실험

속담 인식 실험은 단어 인식에서 인식된 단어열을 바탕으로 Markov 모델을 통하여 만들어진 기준 속담과의 거리를 비교하여 각 속담에 대한 인식을 하게 된다. 4명의 화자에 대한 속담 인식에서는 화자 독립의 인식에서 한개의 속담만이 오인식되어 평균 99.6%의 인식율을 보였다. 속담 인식에서의 인식율의 향상은 속담내에서의 중첩되는 단어가 존재하지 않기 때문에 인식율의 향상을 가져왔다.

V. 결 론

본 논문을 통하여 연속음성에 나타나는 음소와 단어 및 속담을 신경회로망과 Markov 모델을 이용하여 인

식하였다. 25.6ms의 시간창을 사용하여 5ms마다 음성의 특징을 구하였으며, 음성의 시간변화를 학습하기 위하여 300ms의 시간파형에 대하여 선택적으로 입력벡터를 구성하였다. 음성의 특징을 표현하기 위하여 사람의 귀의 특징을 고려한 PLP 모델을 사용하였고, 각 음성에 대하여 2명의 화자의 음성으로부터 학습벡터를 추출하여 학습하고, 학습된 음소열을 이용하여 Markov 모델을 구성하였다. 인식 실험은 화자종속과 독립적인 인식을 위해 각각 화자 2인이 3번씩 발음한 속담을 이용하여 인식을 하였다. 신경회로망을 이용한 음소 인식에서는 자음의 인식율 저하로 평균 81.2%의 음소 인식율을 보였고, Markov 모델을 이용하여 단어 및 속담에 대한 통계값을 구하여 각각의 기준 단어 및 속담을 만들어 인식을 하였다. 인식 결과로 단어 인식에서는 평균 94.0%의 인식율을 보였고, 속담 인식에서는 평균 99.6%의 인식율을 얻었다.

앞으로의 연구 과제로는 대상 어휘나 인식 대상의 수를 늘려 보다 많은 음소와 화자에 대한 인식이 이루어져야 할 것이다.

감사의 글

※ 본 논문은 1995년도 아주대학교 교내연구비 지원에 의해 연구된 것이며, 관계자 여러분께 심심한 감사를 드립니다.

참 고 문 헌

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, pp. 396-453, 1978, Prentice-Hall Inc.

[2] S. Saito and K. Nakata, *Fundamentals of Speech Signal Processing*, 1985, Academic Press.

[3] T. W. Parsons, *Voice and Speech Processing*, pp. 59-81, 1986, McGraw Hill Inc.

[4] P. D. Wasserman, *Neural Computing : Theory and Practice*, 1993, Van Nostrand Reinhold New York.

[5] L. R. Rabinar and Biing-Hwang Juang, *Fundamentals of Speech Recognition*, 1993, Prentice Hall Inc.

[6] H. Hamansky, "Perceptual Linear Predictive(PLP) analysis of speech", *J. Acoust. Soc. Am.*, 87(4) : 1738~1752, April 1990.

[7] Rik D. T. Janssen, Mark Fanty and Ronald, "Speaker Independent Phonetic Classification in Continuous English Letters", *INNS*, Vol. 2, pp. 801~808, 1991.

[8] H. Harmanskey, Kazuhiro Tsuga, Shozo Makino, and Wakita., "Perceptually Based Processing In Automatic Speech Recognition", *ICASSP*, pp. 1971-1162, 1986.

[9] L. R. Rabiner, S. E. Levinson and M. M. Sondhi, "On the Application of Vector Quantization and Hidden Markov Models to Speaker Independent, Isolated Word Recognition", *The Bell System technical Journal*, Vol. 62, No. 4, April 1983.

[10] L. R. Rabiner, J. G. Wilpon and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 37, No. 8, Aug. 1989.

저 자 소 개

洪 基 源(正會員)
1994년 2월 아주대학교 전자공학과(학사). 1994년 3월 ~ 현재 아주대학교 전자공학과 석사과정. 주 관심 분야 : 음성신호처리, 음성인식, 신경회로망

李 幸 世(正會員) 第 32卷 B編 第 11號 參照
현재 아주대학교 전자공학과 교수

金 善 一(正會員) 第 32卷 B編 第 11號 參照
현재 거제전문대 조교수