論文98-35S-8-11

MLP-VQ와 가중 DHMM을 이용한 연결 숫자음 인식에 관한 연구

(A study on the connected-digit recognition using MLP-VQ and Weighted DHMM)

鄭 光 宇 *, 洪 光 錫 **

(Kwang-Woo Chung and Kwang-Seok Hong)

요 약

본 논문에서는 화자 독립 연속 숫자음 인식 시스템의 성능향상을 위하여 MLP-VQ (Multi-Layer Perceptron-Vector Quantizer)를 이용한 가중 DHMM(WDHMM : Weighted Discrete Hidden Markov Models)을 제안한다. MLP 신경망의 출력분포는 입력 패턴과 학습 패턴들간의 비선형 매핑을 통해 각 패턴들간의 유사도를 나타내는 확률분포를 갖는다. 본 논문에서는 MLP 신경망의 출력분포중 가장 높은 출력 값을 갖는 MLP 신경망의 출력 노드의 인덱스를 이용하여 코드워드를 생성하는 MLP-VQ를 제안하였다. 제안된 MLP-VQ는 기존의 VQ에 비해 현재 입력패턴과 학습된 각 class 패턴들간의 유사성 정도를 인식모델에 반영할 수 있는 특징을 갖는다. 또한 MLP 신경망의 출력분포를 DHMM의 심벌 발생 확률의 가중치로 이용하는 가중 DHMM을 구성하였다. 제안된 방법을 이용함으로써 기존의 준 연속 HMM에서 처럼 심벌의 발생확률을 다차원 정규분포로 가정하지 않아도 되기 때문에 HMM 파라메터 추정시간과 인식시간을 단축할 수 있으며, 또한 DHMM보다는 음소 클래스간의 관계를 인식모델에 반영할 수 있기 때문에 적은 계산양의 증가로 인식기의 성능을 14.71%개선할 수 있었다. 실험결과에 의하면, MLP-VQ와 WDHMM에 의한 화자독립 연결 숫자음 인식율은 84.22%이다.

Abstract

The aim of this paper is to propose the method of WDHMM(Weighted DHMM), using the MLP-VQ for the improvement of speaker-independent connect-digit recognition system. MLP neural-network output distribution shows a probability distribution that presents the degree of similarity between each pattern by the non-linear mapping among the input patterns and learning patterns. MLP-VQ is proposed in this paper. It generates codewords by using the output node index which can reach the highest level within MLP neural-network output distribution. Different from the old VQ, the true characteristics of this new MLP-VQ lie in that the degree of similarity between present input patterns and each learned class pattern could be reflected for the recognition model. WDHMM is also proposed. It can use the MLP neural-network output distribution as the way of weighing the symbol generation probability of DHMMs. This newly-suggested method could shorten the time of HMM parameter estimation and recognition. The reason is that it is not necessary to regard symbol generation probability as multi-dimensional normal distribution, as opposed to the old SCHMM. This could also improve the recognition ability by 14.7% higher than DHMM, owing to the increase of small caculation amount. Because it can reflect phone class relations to the recognition model. The result of my research shows that speaker-independent connected-digit recognition, using MLP-VQ and WDHMM, is 84.22%.

* 正會員, 韓國鐵道大學 運轉機電科
(Dept. of Operation-Mechatronics, Korea Railroad College)

** 正會員, 成均館大學校 電氣 電子 및 컴퓨터 工學部
(Dept. of Elec. Eng., Sung Kyun Kwan Univ.)

接受日字: 1996年10月21日, 수정완료일: 1998年6月19日

I. 서 론

음성신호는 비록 언어학적으로 유사한 단어를 발성하더라도 의미, 발성자, 발성속도, 발성자의 심리적, 신체적 상태에 따라 음성신호의 시간적, 주파수적 측면에서 상당히 변질되고 중첩되는 특징을 나타내고 있다[1]. 많은 음성 연구자들이 이러한 음성 변동을 극복할 수 있는 음성 인식 시스템 개발을 위하여 지난 수십 년간 많은 연구를 수행하였다. 그러나 이러한 노력에도 불구하고 음성을 완전하게 인식하는 시스템은 아직 없으며 인식률이 좋다 하더라도 특정 제약조건하에서만 동작되는 특징을 가지고 있다[2-3].

현재 음성인식 시스템에 적용되는 인식 알고리즘으로는 Baker[4]와 Jelinek[5]에 의해 제안된 Hidden Markov Model이 널리 이용되고 있다. HMM은 확률적인 모델로서 음성신호의 시간에 따른 특징변화를 통계적으로 모델링하여 음성을 인식하는 방법이다. 지난 20년간 많은 수학적 이론과 제반 기술들이 발전되어 HMM을 이용한 음성인식이 제한된 분야에서 좋은 인식 성능을 나타내고 있으며, 대용량 화자 독립 음성인식 시스템으로의 확장 면에서 다른 인식 알고리즘에 비해 많은 이점을 갖고 있다. 또한 Cambridge 대학의 HMM Tool Kit[6-7]가 널리 보급되면서 음성 인식 연구자들이 쉽게 음성 인식 시스템을 구성할 수 있게 되었다. 그러나 음성인식의 문제점들이 완전히 해결된 것은 아니며, 많은 부분 보완되어야 하는 점이 남아있다.

이산 분포 HMM의 경우, 음성 신호의 특징을 불연속적인 벡터 코드의 시퀀스로 표현함으로써 음성 분석 시 양자화 오차가 발생되고, 인접 프레임 사이의 상호 관계를 적절히 표현 할 수 없으며, 벡터 코드들간의 상호 관계를 인식 모델에 적용할 수 없는 문제점을 갖고 있다. 이러한 문제점을 부분적으로 극복하기 위하여 준 연속 분포 HMM과 연속 분포 HMM이 제안되어 사용되고 있다. 그러나 이러한 인식 알고리즘은 HMM의 각 상태에서 코드워드의 분포가 가우시안 분포를 갖는다는 가정 하에 벡터 코드워드의 상호 관계를 계산하기 때문에 많은 계산량을 요구하며, 안정된 코드워드 분포를 인식 모델에 구성하기 위해서는 많은 양의 학습 데이터를 요구하게되는 단점이 있으며, 요구되는 수학적 가정들이 음성발성 과정과는 맞지 않다는 문제점이 있다[7]. 또한 기존의 DHMM은 음성 생성 모델에 근거하여 추출한 특징 벡터를 이용하여 VQ를 구성하는데, 단순히 특징 파라메터의 차수간 거리 값에 비례하여 코드워드간의 유사성을 계산하는 것은 음성인지 시스템의 관점에서는 부적합한 표현방법이며, 특징 파라메터의 각 차수가 인지 시스템의 관점에서 미치는 영향 등을 계산하기는 쉽지 않다.

MLP(Multi-Layer Perceptron) 신경망이 제안된 이후 다양한 신경회로망 구조가 개발되어 패턴분류, 시스템 모델링, 신호처리와 같은 많은 분야에서 성공적으로 도입되어 왔다[8]. MLP 신경망은 학습을 통한 입출력 비선형 매핑 기능을 이용하여 음성 신호에 내제된 비선형을 해결하기위하여 제안되었다. 그러나 MLP 신경망의 파라메터가 시간에 대해서 고정되어 있기 때문에 음성과 같은 시변성 신호를 다루기에는 적합하지 못한 단점이 있다. 이러한 단점을 극복할 수 있는 몇 가지 신경망 구조가 제안되었다[9-14].

Waibel등에 의해 제안된 TDNN(Time-Delay Neural Network)은 입력 층과 은닉 층에 음성신호의 시변성을 흡수할 수 있는 구조를 도입하였다[9]. Sakoe등에 의해 제안된 DPNN(Dynamic Programming Neural Network)은 전통적인 동적 프로그래밍 기법과 신경회로망을 접목한 혼합 인식 시스템으로 신경망은 패턴 분류기로 사용되고, 동적 프로그래밍 기법은 음성의 왜곡을 효과적으로 정규화하기 위하여 사용하였다[10]. 그 외에 동적 프로그래밍 기법과 음성의 비선형 예측기로 학습되는 신경망을 결합한 여러 가지 예측신경망이 제안되었으며[11-13], 은닉 뉴런과 입력 뉴런을 연결하는 회귀 연결을 첨가함으로서 음성 신호의 시간적인 변위를 흡수하고 입력신호의 길이에 제한을 두지 않는 회귀신경망(Recurrent Neural Networks:RNN)이 제안되어 음성인식 분야에 적용되고 있다[14].

본 논문에서는 현재 입력 패턴에 대한 각 음성 클래스간의 상관성을 DHMM 인식 시스템에 반영하여 인식기의 성능을 개선하기 위한 방법을 제안한다. 제안된 방법은 MLP (Multi-Layer Perceptron) 신경망의 비선형 매핑을 통한 패턴 분류 기능을 이용하여 MLP-VQ를 구성한다. 이렇게 구성된 신경망의 출력은 입력 특징벡터의 차수별 가중치와 각 음성 클래스간의 상관성을 나타내는 확률분포를 갖는데, 이 확률분포로부터 큰 확률 값을 갖는 출력노드의 인덱스를 이용하여 코드워드를 생성하였다. 이렇게 생성된 코드

워드는 입력 패턴에 대하여 가장 유사성이 높은 각 패턴 클래스간의 관계를 포함하고 있으며, 이 관계를 DHMM에 반영할 수 있는 장점을 갖게된다. 또한 본 논문에서는 신경망의 출력 분포 값을 DHMM 인식 시스템에서 심벌 발생확률의 가중치로 이용하는 가중 DHMM(WDHMM : Weighted DHMM)을 제안하였다.

제안된 인식 모델의 성능 평가를 위하여 연속 숫자음 인식을 수행하였다.

Ⅱ. MLP-VQ

VQ(Vector Quantization)는 음성 분석을 통하여 음성 데이터를 수개의 성분을 갖는 벡터로 재구성하고, 설계된 코드북(codebook)에서 가장 근사한 코드워드를 찾아 그 벡터에 부여된 인덱스를 활용하여 음성 압축 및 인식을 수행한다. VQ의 성능은 코드북 설계에 의해 좌우되며, 코드북 설계 방법으로는 LBG(Linde, Buzo, Gray) 알고리즘이 널리 사용된다[15]. 그러나 LBG 알고리즘은 국소 최적 코드북(locally optimal codebook)만을 보장하기 때문에 초기 코드 벡터 결정이 설계된 코드북 성능에 중요한 영향을 미친다. 또한 코드북 작성후 미지의 입력 벡터와 학습된 임의의 cluster의 중심 벡터 사이에서 Euclidean 거리 값을 계산하여 거리 값이 작은 하나의 cluster만을 선정하기 때문에 두 개의 cluster 경계 면에 존재하는 입력 벡터의 경우에는 많은 오차를 포함하게 되는 단점이 있다.

최근에는 SOFM(Self-Organizing Feature Map)의 경쟁학습 알고리즘을 이용한 VQ가 제안되었다[16]. 그러나 경쟁 학습 알고리즘 또한 학습 전에 학습 벡터 집합이 이루는 초기 cluster의 수를 선정해야 하는 단점이 있으며, 각 cluster에 대한 초기 가중치 벡터를 선택해야 하는 문제점이 있다.

본 논문에서는 MLP 신경망의 학습에 의한 비선형 매핑 기능을 이용하여 VQ 코드를 생성하는 MLP-VQ를 제안한다. MLP-VQ의 코드북 작성은 인식 어휘에 포함된 음소들의 행렬을 이용하여 작성하였다. 표 1은 숫자음에 포함된 음소들에 의해 작성된 코드북을 나타내고 있다.

여기서 각 코드의 인덱스는 MLP 신경망의 출력에 의해 결정되며, 선정 과정은 다음과 같다.

표 1. MLP-VQ의 코드북
Table 1. Codebook of MLP-VQ.

	0	1	2	3	4	5	6	7	8	9	10	11
	12	13	14	15	16	17	18	19	20	21	22	23
	24	25	26	27	28	29	30	31	32	33	34	35
	36	37	38	39	40	41	42	43	44	45	46	47
	48	49	50	51	52	53	54	55	56	57	58	59
	60	61	62	63	64	65	66	67	68	69	70	71
	72	73	74	75	76	77	78	79	80	81	82	83
	84	85	86	87	88	89	90	91	92	93	94	95
	96	97	98	99	100	101	102	103	104	105	106	107
	108	109	110	111	112	113	114	115	116	117	118	119
	120	121	122	123	124	125	126	127	128	129	130	131
	132	133	134	135	136	137	138	139	140	141	142	143

먼저, 현재 입력 패턴에 대한 MLP 신경망의 출력 분포로부터 최대 출력 값을 갖는 두개의 노드를 선정한다.

$$f_{\max} = \mathrm{argmax}(Y_1, Y_2, \ldots, Y_c, \ldots, Y_N)$$
$$s_{\max} = \mathrm{argmax}(Y_1, Y_2, \ldots, Y_c, \ldots, Y_N), \quad s_{\max} \neq f_{\max} \tag{1}$$

만일 $f_{\max} - s_{\max} > \theta$ 이면

$$f_{\max} \leftarrow s_{\max} \tag{2}$$

여기서, N은 출력 음소군의 수이고, Y_c은 MLP 신경망의 출력이고, $f_{\max}$은 가장 큰 출력을 나타내는 출력 음소군의 지수이고, $s_{\max}$은 두 번째로 큰 출력을 나타내는 출력 음소군의 지수이다.

이때 MLP 신경망의 출력분포는 비선형 매핑기능에 의한 입력 패턴과 MLP 학습패턴간의 유사도 를 나타내는 확률분포로 가정할 수 있다. 그중 가장 큰 출력 분포를 갖는 두 개의 노드를 택함으로써 현재 입력 패턴에 대하여 유사도가 높은 패턴 class를 인식 모델에 반영할 수 있는 장점을 지니고 있다.

선정된 두개의 출력 노드를 이용하여 현재 프레임 입력에 대한 유사 코드워드 계산은 다음과 같이 한다.

$$o_m = f_{\max} \times N + s_{\max} \tag{3}$$

여기서, o_m는 현재 입력 특징 벡터에 대한 유사 코드워드이다.

기존의 VQ에서는 각 코드워드에 대한 중심벡터값을 이용하여 코드워드를 발생하는 반면 제안된 MLP-

VQ는 MLP 신경망의 출력분포를 이용하여 유사 코드워드를 발생하기 때문에 각 코드워드에 대한 중심벡터값의 저장이 불필요하고, 새로운 입력 패턴에 대한 거리비교시에 요구되는 계산량이 적은 특징이 있어 숫자음 인식뿐만 아니라 한국어 음성 인식으로 확장이 용이한 특징이 있다.

표 1에는 숫자음 인식의 경우, (식 3)에 의해 얻어진 유사 코드워드의 배열을 나타내고 있다.숫자음의 경우, 각 숫자음의 음소 구조가 5개의 모음과 7개의 자음으로 구성되어 있으며, 각 음소에 대한 분포를 고려하여 신경망을 학습한 후, 입력 패턴에 대한 MLP 신경망의 출력을 (식 3)에 근거하여 가장 큰 첫 번째 출력 음소군의 출력 노드 군의 지수와 두 번째로 큰 출력 음소군의 출력 노드 군의 지수를 이용하여 표 1과 같은 유사 코드워드를 발생 할 수 있다.

표 2와 3에서는 숫자음 인식시 에러 발생 확률이 높은 유사 음절 사이의 유사 코드워드의 분포 예를 나타내고 있다. 표 2는 /일/에 대한 MLP 출력의 유사 코드워드의 분포를 나타내고 있으며, 표 3은 /이/에 대한 MLP 출력의 유사 코드워드의 예를 나타내고 있다. 두 음절 사이의 유사 코드워드의 분포를 살펴보면, 숫자음 /일/의 경우 대다수의 입력 특징 벡터의 유사 코드워드가 22(1st 출력 - /이/ , 2nd 출력 - /ㄹ/인 경우)에서 발생됨을 알 수 있으며, 숫자음 /이/의 경우, 입력 특징 벡터의 유사 코드워드가 20(1st 출력 - /이/, 2nd 출력 - /ㅊ/인 경우)와 23(1st 출력 - /이/ , 2nd 출력 - /ㅁ/인 경우)에서 발생됨을 알 수 있다.

표 2. 숫자음 /일/에 대한 MLP 출력의 유사 코드워드의 분포

Table 2. Similar codeword distribution of MLP outputs for the digit /1/.

	28		1	11		28		20	14		58
				1		2					
	5					7				5	
	1			1						1	
	1										
			1	8		2				3	4
	29			2						5	

표 3. 숫자음 /이/에 대한 MLP 출력의 유사 코드워드의 분포

Table 3. Similar codeword distribution of MLP outputs for the digit /2/.

	10				5			40	35	
	2				1			1		
	1									
	10									

이는 두 번째 출력 음소 군까지 고려함으로써 입력 특징 벡터에 최대 유사도를 갖는 음소뿐만 아니라 학습 음소간의 유사도를 두 번째 출력 음소군 노드에 반영함으로써 서로 유사한 음절간의 변별 능력을 개선할 수 있다.

Ⅲ. 가중 이산 분포 HMM

이산 분포 HMM 방법은 입력으로 사용되는 벡터 양자화가 연속적인 음성의 특징량을 이산화하기 때문에 양자화 왜곡이 발생하고, 다른 코드워드와의 관계를 표현하기에는 부적합한 단점이 있으며[17], 연속 분포 HMM방법은 양자화 왜곡을 없앨 수 있으며, 입력 벡터와 여러 코드워드와의 관계를 모델에 반영할 수 있다는 장점을 가지고 있으나 계산량이 매우 많아지는 단점이 있다[18]. 기존의 단어 모델에 기반을 둔 HMM은 인식 단어의 수가 증가함에 따라 선형적으로 비례하여 HMM 모델 수가 증가 하므로 대어휘 연속 음성 인식 시스템에는 적합하지 않다. 그러므로 보다 일반적으로 적용할 수 있는 인식단위가 필요하다. 이러한 인식단위로는 음소, 음절, diphone등과 같은 부단어 단위가 있으며, 이러한 부단어 단위를 HMM 모델에 적용하기 위한 방법들에 대하여 많은 연구가 진행되고 있다.

일반적인 준 연속 분포 HMM은 적은 계산량과 양자화 왜곡을 최소화하고, 각 입력 벡터에 대한 각 코드워드의 출력 확률은 각 코드워드를 평균적으로 하는

다차원 정규 분포에 대한 입력 함수 값과 학습에 기초한 HMM 파라메터 추정에 의해 얻어지는 각 코드워드에 대한 심벌 출력 확률과의 선형 결합으로 주어진다[19]. 각 코드워드를 평균 벡터로 하는 다차원 정규 분포의 공분산 행렬은 HMM 파라메터의 추정 시에 각 단어에서 독립이고 각 모델 내에서는 전체 상태 천이에 공통인 것으로 하여 구한다. 일반적으로 준 연속 HMM은 DHMM을 각 코드워드의 출력 확률에 대한 보간 계수의 도입에 의해서 확장한 것, 또는 혼합 정규 분포에 의한 연속 분포 HMM에서 평균 벡터 및 공분산 행렬을 공동화한 간략한 방법으로 생각할 수 있다.

본 논문에서는 MLP 신경망의 출력 분포가 입력벡터와 학습벡터간의 유사도에 따른 확률분포로 가정하고, 이 출력 값을 DHMM 모델의 각 코드워드에 대한 출력 확률의 보간 계수로 이용하는 가중 이산 분포 HMM(WDHMM:Weighted DHMM) 모델을 제안한다. 제안된 방법에서는 각 입력 벡터에 대한 출력 확률 계산시 기존의 준 연속 분포 HMM 모델의 경우에서처럼 각 코드워드를 평균 벡터로 하는 다차원 정규 분포에 대한 함수로 가정하여 계산하는 과정이 불필요하고, 단지 MLP 신경망의 출력 값을 직접 이용하기 때문에 입력벡터에 대한 출력 확률계산이 아주 간단한 특징을 갖고 있으며, 입력 벡터와 학습 벡터간의 유사도를 DHMM 모델에 반영할 수 있다.

제안된 WDHMM은 두 단계로 구분할 수 있다. 첫 번째는 HMM의 학습단계로 MLP-VQ에서 발생한 코드워드를 이용하여 HMM을 학습한다. 이는 기존의 VQ에서 발생한 코드워드대신 MLP-VQ에 발생한 유사 코드워드를 이용하는 것을 제외하고는 기존의 DHMM을 학습하는 방법과 동일하다. 두 번째 단계는 인식단계로 MLP 신경망의 출력 분포를 DHMM의 심벌 발생 확률의 가중치로 이용한 단계이다. 이는 학습 단계에서 얻은 HMM 모델의 심벌 발생 확률에 MLP 신경망의 출력을 보간 계수로 이용함으로써 미지의 입력 패턴과 학습 패턴과의 유사도를 인식 모델에 반영할 수 있는 특징을 갖는다.

WDHMM의 인식단계에서 심벌 발생 확률은 다음과 같이 구해진다

$$W_{ij}(o_m) = Y_{f_{max}} \times b_{ij}(o_m) + Y_{s_{max}} \times b_{ij}(o_m) \quad (4)$$

여기서 $b_{ij}(o_m)$은 DHMM의 심벌 발생 확률이고, Y_j은 MLP 신경망의 출력이다.

표 4에는 인식과정에서 각 HMM에서 출력 확률을 산출하는 방법을 비교하여 나타내었다.

표 4. 각 HMM에서 출력 확률 산출방법
Table 4. Computation of output probabilities for each HMM.

	출력확률 산출방법	보간방법
	$b_{ij}(o_k): k = \arg nin\, d(X_t, o_m)$	없 음
	$W_{ij}(X_t) = \sum_m N(X_t, \mu_m, \sum_m) \times b_{ij}(o_m)$	다차원 정규분포
	$W_{ij}(o_m) = Y_{f_{max}} \times b_{ij}(o_m) + Y_{s_{max}} \times b_{ij}(o_m)$	MLP 신경망의 출력

위 과정을 통하여 구한 입력 벡터에 대한 유사 코드워드의 MLP 출력을 DHMM의 $b_{ij}(o_m)$에 보간 계수로 이용함으로써 간단한 계산 과정을 통하여 WDHMM으로 확장할 수 있으며, 유사 음소간의 관계와 유사성 정도를 DHMM내에 반영할 수 있어 성능을 개선할 수 있다.

Ⅳ. 실험 및 고찰

1. 실험 조건

본 논문에서는 제안된 인식 모델과 알고리즘을 평가하기 위하여 연결 숫자음(1~4자리 숫자열)을 대상으로 실험하였다. 실험에 사용된 데이터는 13명의 화자가 140개의 숫자열을 6회 발성한 음성을 이용하였다.

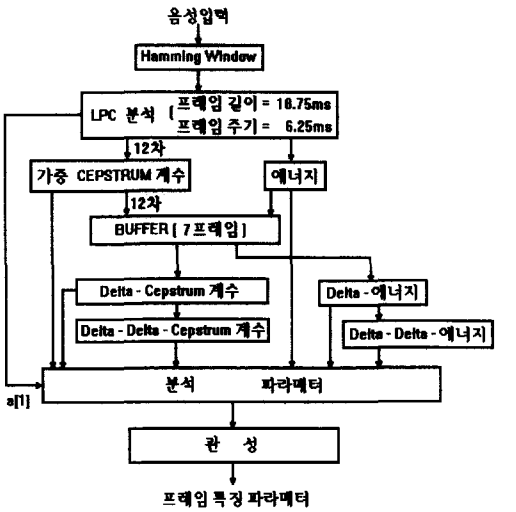


그림 1. 특징 파라메터 추출
Fig. 1. Feature parameter extraction.

표 5. 분석 조건
Table 5. Analysis conditions.

샘플링 주파수	16 kHz
분석 프레임 길이	18.75 ms
프레임 주기	6.25 ms
1차 LPC 계수	1 차
영교차율	1 차
에너지	1 차
Delta-energy	1 차
Delta-delta-energy	1 차
가중 LPC cepstrum	9 차
Delta-cepstrum	9 차
Delta-delta-cepstrum	6 차

음성 데이터에 대한 특징 파라메터 추출과 분석 조건은 그림 1과 표 5에 나타내었다.

입력 특징벡터의 차수는 12차 가중 LPC Cepstrum 계수[20]와 Delta-cepstrum 계수[21-22], Delta-delta-cepstrum 계수[23] 중 각 계수가 인식률에 미치는 영향 등을 고려하여 가중 LPC Cepstrum과 Delta-cepstrum은 9차까지 이용하였으며, Delta-delta-cepstrum은 6차까지 이용하였다[20]. 또한 에너지와 영교차율 등을 포함한 5개의 파라메터를 결합하여 총 29개의 특징벡터로 구성하였다.

음성 신호에 대한 특징 벡터 시퀀스의 안정성은 음성 인식 시스템의 성능에 많은 영향을 주게되는데, 대개의 경우 특징 벡터의 시퀀스에 대한 벡터 공간상의 궤적은 국소구간에서 랜덤한 변동을 나타내고, 이는 특징 벡터 공간상에서 특징 벡터들의 확률 분포을 증첩시키는 하나의 요인이 된다[24]. 이러한 특징 벡터 시퀀스 상에서 발생하는 랜덤한 변동을 제거하고, 궤적을 안정화 시키기 위하여 자기회귀(autoregressive) 필터를 사용하여 특징 벡터의 시퀀스에 대한 궤적을 smoothing 처리하였다.

$$\vec{Y}(t) = a \cdot \vec{X}(t) + b \cdot \vec{Y}(t-1) + c \cdot \vec{Y}(t-2) \quad (5)$$

특징 벡터 시퀀스에 관성을 주기 위해 사용한 필터의 계수 값은 음소의 평균 길이를 고려하여 적절하게 선택해야 하는데, 필터 계수에 대해 a=0.25 , b=0.72 , c=0.1로 각각 선택하였다. 이러한 필터 계수의 선택은 반복적인 실험을 통한 결과 값으로, 필터의 탭수가 작을 경우 특징 벡터 시퀀스상의 안정화를 볼 수 없었으며, 반면 필터의 탭수가 너무 크면, 특히 짧은 음절에 대해 음소간 천이가 불분명해져 인식률이 오히려 감소되었다[24]. 그러므로 필터 계수는 주어진 분석 조건에 대해 특징 벡터 시퀀스 상의 안정화와 음소간 천이 특성을 동시에 고려하여 적절히 선택하였다.

그림 2에는 본 실험에서 사용된 MLP 신경망의 구조를 나타내고 있다. 실험에 사용된 MLP 신경망은 학습 속도의 개선을 위하여 모음과 자음을 인식하기 위한 신경망과 유성음과 무성음을 구분하기 위한 신경망등 3개의 부분망을 구성하였다.

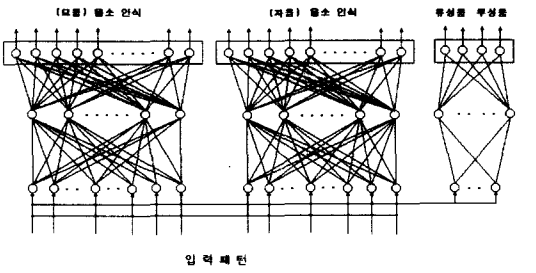


그림 2. 실험에 사용된 MLP의 구조
Fig. 2. Structure of the MLP used in experiments.

2. 연결 숫자음 인식

본 실험에서는 1자리 숫자음에서부터 4자리 숫자음으로 구성된 140개의 연결 숫자음을 선정하여 인식을 수행하였다. 140개의 데이터 목록중에서 4연속 숫자음은 조음현상을 고려하여 ETRI에서 선정한 35개를 이용하였으며, 이를 기반으로 모음연쇄와 조음특성 등을 고려하여 단음절 숫자음에서 부터 3연속 숫자음까지 총 105개를 추출하여 사용하였다. 실험에 사용된 데이터는 13명의 화자가 6회씩 발성한 데이터를 이용하였다. MLP 신경망의 학습과 가중 DHMM 모델을 작성하기 위하여 5명의 화자가 발성한 데이터 중 첫번째 발성한 음성 데이터를 두개의 그룹으로 나누어 학습 데이터를 이용하였다.

첫번째 그룹의 학습데이터는 MLP 신경망 학습을 위하여 이용하였으며, 신경망 학습을 위한 데이터는 각 숫자음에 포함된 음소가 고르게 분포하도록 선정하였다. 또한 MLP 신경망의 학습에 사용된 데이터와 그 외의 학습 데이터를 이용하여 가중 DHMM 모델을 작성하였다.

나머지 데이터는 화자종속 및 화자 독립 실험을 위해서 사용하였다.

또한 실험에 사용된 연결 숫자음의 목록은 표 6에

나타내었다.

표 6. 음성 데이터 목록
Table 6. List of the speech data.

0287	0316	0721	1199	1398	1427	1823	2244
2409	2538	2934	3045	3510	3649	4156	4621
4750	5267	5500	5737	6378	6633	6843	6872
7083	7489	7954	8065	8194	8590	8877	9176
9205	9601	9861	028	031	045	065	072
119	139	142	156	176	182	199	224
240	267	253	287	304	316	378	409
427	462	475	500	510	526	538	550
573	590	601	621	633	649	663	684
687	708	737	748	750	795	806	819
823	845	859	861	872	877	887	917
920	954	00	02	05	06	07	08
09	11	13	15	16	19	20	21
22	24	25	30	35	36	37	39
40	42	44	45	50	51	56	60
61	62	77	72	79	82	85	92
96	99	0	1	2	3	4	5
6	7	8	9				

연결 숫자음 인식 시스템의 전체 구성은 그림 3과 같다.

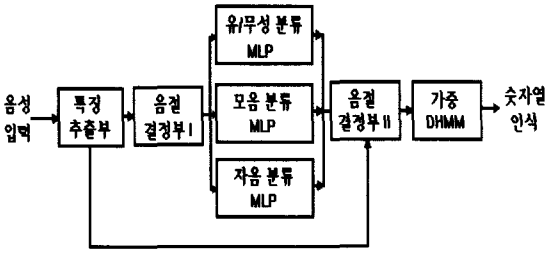


그림 3. 연결 숫자음 인식 시스템의 구성도
Fig. 3. Block diagram of the connected-digit recognition system.

본 논문에서 사용한 연결 숫자음 인식을 위한 인식 시스템은 크게 음절 결정부 I, 음소 분류 신경망, 음절 결정부 II, 음절 인식부등 네 부분으로 나눌 수 있다.

음절 결정부 I은 음성신호의 프레임별 에너지와 음절의 길이 정보를 이용하여 연결 숫자음 중에서 묵음 구간에 의해 분리되어 발성되는 음절 영역 구간을 검출하기 위하여 사용된다. 여기서 사용된 에너지는 특징 벡터 추출시 음성 구간의 최대 값으로 현재 프레임을 정규화시킨 정규화된 에너지를 이용하였다.

음소군 분류 신경망은 음절 결정부 I에서 세그멘테이션된 음절 영역 부분의 특징 파라메터를 표 7에 나타낸 음소군으로 분류하는데 이용하였다. 또한 음소 smoothing을 이용하여 1~3 프레임의 불연속적인 변동을 하나의 음소로 처리하였다. 이러한 처리를 통하여 입력 음절 영역에 대하여 대략적인 음소군 분류를 수행한 후, 그 처리 결과를 음절 결정부 II로 넘겨준다.

표 7. 음소군 분류
Table 7. The classification table of phoneme groups.

음소 / 그룹	음 소 군
그 룹 I	ㅏ
그 룹 II	ㅣ
그 룹 III	ㅗ ㅜ
그 룹 IV	육
그 룹 V	ㄹ ㅁ ㅇ
그 룹 VI	ㅅ ㅊ ㄱ ㅍ

음절 결정부 II는 음절 결정부 I에서 분리된 음절영역 정보로 부터 연음되어 발생된 음절을 분리하고, 연결 숫자음 내에 포함된 최종적인 음절수를 결정하기 위한 단계로서 음소군 분류 신경망의 출력 값과 입력 음절의 시간 길이 정보, 에너지의 dip등을 이용하여 네 가지의 결정규칙을 설정하여 연음되어 발성된 음절 영역으로 부터 숫자음 음절을 분리하도록 하였다.

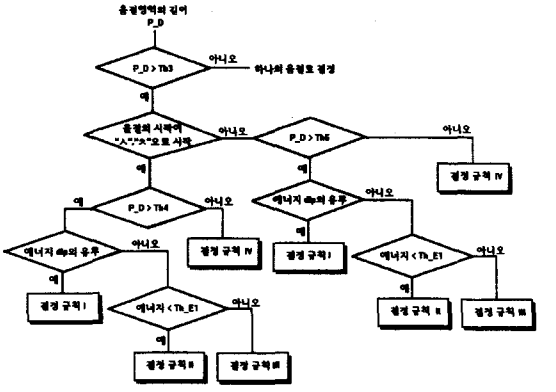


그림 4. 음절 결정부 II의 알고리즘
Fig. 4. Algorithm for the syllable decision part II.

결정 규칙 I - 두음절 이상의 음절 영역 안에 에너지 dip이 발생하는 경우

1) 에너지 dip 부분을 초기 음절 분리 프레임으로 설정
2) 에너지 dip 부분의 전반부와 후반부의 음소군 분

류 결과를 조사하여 불연속적인 모음열이나 유성 자음이 존재하는 부분을 음절 경계로 설정하여 초기 음절 분리 프레임을 조정한다.

3) 분리된 음절의 길이가 DR_1 프레임보다 적으면 분리된 음절을 하나의 음절로 결합한다.

여기서, 사용된 에너지 dip 부분은 음절영역의 초성이 파열음인 경우에 음절의 앞부분에서 발생 될 수 있으므로, 초성부분에서 발생되는 에너지 dip 부분을 음절의 경계로 잘못 선택되지 않도록 파열음의 최대 길이보다 큰값을 임계치(DR_1=23)로 사용하여 잘못된 에너지 dip을 수정하였다.

결정규칙 II

- 두음절 이상의 음절영역에서 후속 음절의 초성 자음이 무성자음으로 시작되는 경우

1) 음절 영역의 중앙부분을 조사하여 초성 자음의 평균에너지 Th_E1보다 작은 영역을 초기 음절 경계로 설정한다.
2) 초기 음절 경계로부터 초성 자음의 시작부분을 조사하여 초기 음절경계를 수정한다.
3) 분리된 음절의 길이가 DR_1 프레임보다 작으면 분리된 음절을 하나의 음절로 결합한다.

여기서 Th_E1은 무성자음의 평균에너지를 이용하여 0.2로 설정하였다.

결정규칙 III - 모음이 연속되는 경우와 종성자음과 모음이 연속되는 경우

1) 음절 영역의 중앙을 초기 음절 분리 프레임으로 설정한다.
2) 초기 음절 분리 영역으로부터 음소군 분류 신경망의 출력을 이용하여 종성 자음이 존재하는 부분 혹은 모음열이 불연속적인 구간을 조사하여 초기 음절 분리 영역을 수정한다.
3) 동일한 모음이 연속인 경우는 음절 영역의 길이를 조사하여 두 음절 혹은 세 음절로 균일하게 분할한다.

결정규칙 IV - 음절 영역이 하나의 음절로 구성되어 있는 경우

1) 음절 영역을 하나의 음절로 처리한다.

3. 연결 숫자음 인식 실험

제안된 방법의 타당성을 검증하기 위하여 제안된 방법을 개별적으로 적용하여 실험하였다.

연결 숫자음 인식에 관한 실험 결과를 표 8에 나타내었다. 여기서 사용된 실험데이터는 표 6에 기술한 140개의 목록에 포함된 연결 숫자음을 이용하였다. 연결 숫자음 인식 실험의 결과를 고찰해 보면, 방법 I과 방법 IV은 관성항이 인식 시스템에 미치는 영향을 조사하기 위한 실험이다. 방법 I과 방법 IV을 비교해 보면 관성항을 적용함으로써 화자 종속의 경우 4.5%의 인식률 향상과 화자 독립의 경우 4.88%의 인식률 향상을 얻을 수 있었다. 이는 인접프레임 사이에 자기회귀 필터를 사용함으로써 벡터 시퀀스상에 발생하는 랜덤한 변동을 제거하고, 궤적을 안정화 시켜, 벡터 공간상에서 특징벡터들의 확률분포가 중첩되는 것을 억제하기 때문에 좋은 성능을 얻을 수 있었다.

표 8. 연결 숫자음 인식에 관한 실험 결과
Table 8. Result of the connected digits recognition experiments.

방 법 \ 종속/독립	화자 종속 실험	화자 독립 실험
관성을 적용하지 않은 경우(방법 I) ***MLP-VQ, WDHMM***	83.16 %	79.34 %
이산 HMM 모델의 경우(방법 II) ***관성, MLP-VQ, DHMM***	77.49 %	69.51 %
SCHMM을 이용한 경우(방법 III) ***관성, MLP-VQ, SCHMM***	83.29 %	73.78 %
제안된 방법을 이용한 경우(방법 IV) ***관성, MLP-VQ, WDHMM***	87.66 %	84.22 %

방법 II와 방법 IV은 DHMM 모델인 경우와 가중 이산 분포 HMM 모델의 성능 평가를 위한 실험이다. 이 경우 제안된 방법이 화자 종속의 경우 10.17%, 화자 독립의 경우 14.71%의 인식률 향상을 나타내고 있다. 이 경우, 방법 II의 인식률이 상당히 낮게 나타나고 있는데, 이는 DHMM 모델을 이용하여 실험할 경우 음소사이의 유사도를 인식단계에서 이용할 수 없으므로 유사 음절 사이에서 오인식이 많이 발생하였기 때문이다.

방법 III은 본 논문에서 제안한 가중 DHMM을 보는 관점에 따라 기존의 DHMM에 MLP 신경망의 출력을 가중치로 이용한 방법으로 볼 수 있으며, 또한 기존의 SCHMM에서 다차원 정규분포에 의해 구해지는 보간계수를 신경망의 출력값으로 대치하여 간략화한 것으로 볼 수 있다. 이런 두가지 관점에 대한 평가를 비교하기 위하여 MLP-VQ와 SCHMM을 이용한 연결

숫자음 인식 실험을 수행하였다. 그러나 제안된 MLP-VQ는 기존의 VQ처럼 각 코드워드에 대한 중심벡터를 갖고 있지 않기 때문에 본 실험에서는 MLP-VQ에 의해 구해진 각 코드워드를 기초로 각 코드워드에 속하는 학습 데이타를 별도로 수집하여 각 코드워드에 대한 중심벡터와 분산을 구하여 SCHMM에 이용하였다. 실험결과에 의하면 MLP-VQ와 SCHMM을 이용한것보다 제안된 방법(IV)이 더 좋은 성능을 나타내고 있음을 알 수 있다. 이는 비선형 매핑 기능을 이용하는 MLP-VQ에 의해 생성된 VQ 코드워드들의 중심벡터와 SCHMM의 보간 계수를 구하기 위하여 별도로 구한 VQ 코드워드의 중심벡터간에 불일치로 인하여 MLP-VQ와 SCHMM을 이용한 방법의 성능이 저하되는 것으로 생각된다.

방법 IV는 본 논문에서 제안한 WDHMM 인식 모델을 이용하여 연결 숫자음을 인식한 경우로 화자 종속의 경우 87.66%의 인식률을 얻을 수 있었으며, 화자 독립의 경우 84.22%의 인식률을 나타내고 있다.

V. 결 론

본 논문에서는 화자 독립 음성 인식 시스템의 성능을 개선하기 위하여 새로운 형태의 인식 모델과 알고리즘을 제안하여 검토하였다. 다양한 형태의 음운 환경에서 발성되는 음소의 특징 분포를 신경망의 학습 기능을 이용하여 인식 시스템 내에 표현하기 위하여 MLP-VQ 구조를 제안하였다. 또한 MLP 신경망의 출력 값이 입력 음성 패턴과 학습 음성 패턴과의 유사도를 나타내는 지표로 이용하여 새로운 형태의 가중 DHMM 모델과 파라메터를 추정하는 방법을 제안하였다. 또한 입력 특징 벡터 시퀀스의 변동을 줄이고 특징 벡터 시퀀스상에서 인접 프레임간의 상관성을 고려하기 위하여 관성항을 도입하였다.

제안된 알고리즘들은 혼합 인식 알고리즘에 기초한 화자 독립 인식 시스템의 성능을 개선하는 방법들이며, 연결 숫자음 인식을 통해 제안된 인식 알고리즘의 성능을 검토하였다.

입력 음성과 학습 음성과의 유사도에 기초한 가중 DHMM 모델은 MLP 신경망의 출력이 입력 음성과 학습 패턴과의 유사도에 따른 확률 분포로 가정하고, MLP 신경망의 출력 값을 DHMM모델 파라메터 추정과 인식시 심볼발생확률의 보간 계수로 이용하는 방법이다. 또한 HMM 파라메터의 추정시간을 단축하기 위하여 MLP 신경망의 출력값 중 가장 큰 두개의 출력 음소를 추출하고, 각 음소에 대한 출력 노드군의 지수를 코드워드로 사용하여 유사 코드워드를 생성하는 방법을 제안함으로서 유사 음소간의 관계 및 현재 입력 패턴과 학습 패턴 사이의 유사도 정도를 인식 모델에 적용하는 방법을 제안하였다. 제안된 가중 DHMM 모델은 기존의 SCHMM 모델처럼 각 코드워드가 다차원 정규 분포로 가정하지 않아도 되기 때문에 SCHMM의 파라메터 추정 시간과 인식 시간을 단축할 수 있었으며, DHMM보다는 적은 계산량의 증가로 인식 시스템의 성능을 최대14.7% 개선할 수 있었다. 그러나 부정확한 세그멘테이션에 의해 발생되는 에러가 전체 인식 시스템에 미치는 영향이 상당히 크게 나타나고 있는데, 인식 시스템의 성능 개선을 위해서는 좀더 정확한 세그멘테이션 규칙이 필요할 것으로 판단된다.

참 고 문 헌

[1] S.Furui, "Speaker-independent and speaker-adaptive recognition techniques", *Advances in Speech Signal Processing*, pp.597-622, Marcel Dekker Inc., 1992.

[2] W.A.Lea, *Trands in speech recognition*, Englewood Cliffs, NJ:Prentice-Hall, 1980.

[3] J.Mariani, "Recent advances in speech processing", Proc.ICASSP-89, pp.429-440, 1989.

[4] J.Baker, "The DRAGON System - An overview", IEEE ASSP, vol.23, no.1, pp.24-29, 1975.

[5] F.Jelinek, "Continuous speech recognition by statistical methods", Proc.IEEE, vol.64, no.4, pp.532-555, 1976.

[6] P.Woodland, and S.Young, "The HTK tied-state continuous speech recognizer", Eurospeech'93, pp.2207-2210, 1993.

[7] N.Morgan, and H.Bourlard, "Continuous Speech Recognition", IEEE Signal Processing Magazine, pp.25-42, May 1995.

[8] W.Huang, R.Lippmann, and T.Nguyen, "Neural Nets for Speech Recognition", Conf. of the Acoustic Society of America, Seattle WA, 1988.

[9] A.Waibel, T.Hanazawa, G.Hinton, K. Shikano and K.Lang, "Phoneme recognition using time-delay neural networks", Proc.ICASSP'88, pp.107-110, 1988.

[10] H.Sakoe, R.Isotani, K.Yoshida, K.Iso and T.Watanabe, "Speaker-independent word recognition using dynamic programming neural networks", Proc.ICASSP'89, pp. 29-32, 1989.

[11] E.Levin, "Word recognition using hidden control neural architecture", Proc. ICASSP'90, pp.433-436, 1990.

[12] K.Iso and T.Watanabe, "Speaker-independent word recognition using a neural prediction model", Proc.ICASSP'90, pp.441-444, 1990.

[13] J.Tebelskis and A.Waibel, "Large vocabulary recognition using linked predictive neural networks", Proc. ICASSP '90, pp.437-440, 1990.

[14] R.J.Williams and D.Zipser, "A learning algorithm for continually running fully recurrent neural networks", Neural Computation, vol.1, pp.270-280, 1989.

[15] Y.Linde, A.Buzo, and R.M.Gray, "An algorithm for vector quantizer design", IEEE Trans. Com., vol.28, pp.84-95, 1980.

[16] S.C.Ahalt, A.K.Krishnamurthy, P.Chen, and D.E.Melton, "Competitive learning algorithms for vector quantization", Neural Networks, vol.3, pp.277-289, 1990.

[17] L.R.Rabiner, "A Tutorial on hidden Markov models and selected applications in speech recognition", Proc.IEEE, pp. 257-268, Feb., 1989.

[18] L.R.Rabiner et al, "Some properties of continuous hidden Markov model representation", AT&T Tech.J., vol 64, pp.1251-1270, July-Aug., 1985.

[19] X.D.Huang, Y.Ariki, M.A.Jack, *Hidden Markov Model for Speech Recognition*, Edinburgh University Press, 1990.

[20] E.L.Bocchieri, J.G.Wilpon, "Discriminative feature selection for speech recognition", Computer speech and language vol. 7, pp.229-246, 1993.

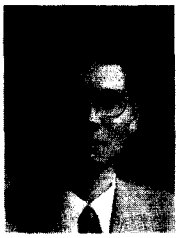
[21] S.Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE ASSP, vol.34, No.1, pp.52-59, 1986.

[22] B.A.Hanson, T.H.Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration features : experiments with Lombard and noise speech", Proc.ICASSP'90, pp.857-860, 1990.

[23] K.F.Lee, H.W.Hon, "Speaker independent phone recognition using hidden Markov models", IEEE ASSP, vol.37, pp.1641-1648, 1989.

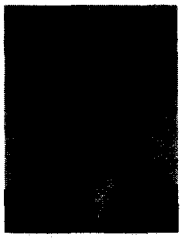
[24] 정 광우, 윤 석현, 홍 광석, 박병철, "관성과 SOFM-HMM을 이용한 고립단어 인식," 대한전자 공학회 논문지, 제 131권 B편, 제6호, pp.716-723, 1994

저 자 소 개



鄭 光 宇(正會員)

1989년 성균관대학교 전자공학과 졸업. 1991년 성균관대학교 전자공학과 공학석사. 1995년 성균관대학교 전자공학과 공학박사. 1996년 3월 ~ 현재 한국철도대학 운전기전과 전임강사. 주관심분야는 음성 및 신호처리, 철도 신호 및 자동화 시스템



洪 光 錫(正會員)

1985년 성균관대학교 전자공학과 졸업. 1988년 성균관대학교 전자공학과 공학석사. 1992년 성균관대학교 전자공학과 공학박사. 1990년 3월 ~ 1993년 2월 서울보건전문대학 전산정보처리과 전임강사. 1993년 3월 ~ 1995년 2월 제주대학교 정보공학과 전임강사. 1995년 3월 ~ 1996년 2월 성균관대학교 전자공학과 조교수. 1996년 3월 ~ 현재 성균관대학교 전기·전자 및 컴퓨터공학부 조교수. 주관심분야는 음성 및 신호처리, HCI