음소단위 코드북간의 확률적 전이 모델을 이용한 한국어 숫자음 인식에 관한 연구

최 환 진*, 오 영 환
한국과학기술원 전산학과

Isolated Korean Digits Recognition Using Stochasitc Transition Models with Phoneme-based VQ Codebooks

Choi Hwan Jin*, Yung Hwan Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology

요 약

음성인식을 위해 다양한 방법들이 제안되어 있다. 본 연구에서는 음소단위 각각의 벡터 양자화된 코드북의 색인를 학습하는 HMM을 이용하여 한국어 숫자음을 대상으로 인식 실험을 수행하였다. 실험결과, 기존의 단어단위 HMM과 음소단위로 이루어진 유한상태기계(FSM)구조의 인식기에 비해 높은 인식율을 보였다.

1 서론

음성은 인간의 언어소통을 위한 가장 기본적인 수단으로 사용의 간편함이나 의사 전달의 효율성등의 측면에서 다른 방법들에 비해 우수한 것으로 알려져 있다. 인간과 기계간의 의사소통을 위해서 인간이 발성한 음성을 입력으로 받아서 음향처리와 인식과정을 거쳐 음성 해독을 수행하는 음성인식에 관한 다양한 연구가 현재 진행되고 있는 실정이다.

음성인식을 위해 사용되는 방법은 크게 두 가지 방향으로 나눌 수 있다[1]. 첫번째 방법으로 인간의 인식방법을 지식 공학적인 측면에서 모델링하는 방법이 있으며, 두번째로 패턴인식에 기반을 둔 방법이 있다. 지식공학에 기반을 둔 방법의 경우 언어적으로 정의되는 단위 이외에 다양한 음성환경에 따른 인식단위의 변이를 규칙의 형태로 표현하여 인식하는 방법으로, 표현규칙의 제약성, 많은 기억장소의 요구 및 일관성 있는 추론등의 난점들을 가지고 있다. 반면, 패턴인식에 기반을 둔 방법의 경우 동

일한 분류에 속하는 자료의 특성을 이용하여 인식을 수행한다. 음성과 같이 변이가 큰 패턴의 경우 수집된 자료의 분석을 통한 모델링 보다는 수집된 음성자료와 언어적으로 정의되는 단위 간의 사상관계를 인식 시스템이 자동적으로 추출 하므로써, 시스템의 구현이 비교적 단순하며, 인식율이나 성능면에서도 우수한 것으로 알려져 있다. 음성인식을 위해 신경회로망[1], 벡터 양자화(VQ)[5], hidden Markov Model(HMM)[2]등의 방법이 사용되고 있다.

본 연구에서는 음소형태로 분할된 음성자료를 사용하여 VQ와 HMM을 결합한 형태의 시스템을 제안한다. VQ를 이용한 대표적인 음성인식방법으로는 인식대상 음성을 시간적으로 정규화하고 정규화된 시간을 등간격으로 구분한 후 각각에 대해서 개별 코드북을 작성하는 Burton이 제안한 multisection 코드북작성 방법이 있다[6]. 이 방법의 경우 음성에 내포된 시간적인 특성을 반영하고 있으나, 시간에 대한 정규화에 따라 지속시간의 차이로 동일한 음성이 다른 시간구간에 포함될 수 있으며, 화자 독립의 경우 인식율이 낮은 단점을 갖고 있는 것으로 알려져 있다.

이러한 방법과는 달리, 인식 대상어휘를 subword단위로 나누고 각각에 대해서 유한상태기계(finite state machine ; FSM)구조로 구성하여 인식을 수행하는 방법이 있다. Huang[7]나 Kopec[8]에 의해서 제안된 이 방법은 인식단어간에 공유되는 음소를 하나의 코드북으로 구성하므로써 위치상으로 서로 다른 동일한 음성에서 발생되는 오류를 줄일 수 있으며, 아울러 길이가 짧은 파열음이나 마찰음의 경우 단일 코드북사용시 제외될 가능성이 크나 별도의 코드북을 사용하므로써 자음에 대해 오류를 감소시킬 수 있다. 그러나, Huang이 제안한 FSM 형식의 단어 모델을 구성하는 경우 상태(state)간의 전이는 입력자료에 대해서 인접상태에서 각각의 해당 코드워드에 대한 오류를 최소화하는 방향으로 전이가 이루어진다. 일단 다른 상태로의 전이가 이루어지면 이전 상태로의 전이가 불가능하다. 이러한 경우 상태간의 전이가 불안정하게 이루어질 가능성이 존재한다. 즉, 음성신호에서의 약간의 변화가 인식시 큰 오차를 유발할 수 있으며, 이러한 결과는 부적절한 천이로 인해 발생될 가능성이 크므로 Huang이 제안한 천이 방법보다는 좀더 신뢰성이 있는 방법이 요구된다. 따라서, 음식대상어휘를 보다 신뢰성을 가지고 인식하기 위해서 음소단위로 분할된 학습자료를 사용한 확률적인 전이 특성을 갖는 모델을 본 논문에서 제안하고자 한다.

본 연구에서 제안된 모델은 음소 코드북의 색인의 열을 학습한 HMM과 학습된 HMM의 각상태로부터 추정된 각 코드북의 코드워드의 존재 확률의 2 가지로 구성된 모델을 이용하여 인식을 수행한다. 다음의 2 절과 3 절에서 제안된 인식 시스템의 구조와 기능, 인식과정등에 대해서 상세히 기술하며, 4절에서는 제안된 시스템과 HMM, FSM구조의 인식기간의 성능비교 실험 및 결과를 분석하며, 마지막 5절에서 결론을 맺고자 한다.

2 시스템 개요 및 구성

음성인식을 위해서는 일반적으로 다단계의 처리과정이 필요하다. 이러한 처리과정은 크게 학습과정과 실험과정으로 분류할 수 있으며, 학습과정은 다시 음성 자료 수집단계, 음향처리 및 분석 단계, 인식단계등으로 세분된다. 이러한 2단계 처리과정에서 첫번째 단계로부터 얻어진 결과가 두번째 단계에 미치는 영향이 크므로, 인식 대상에 알맞는 인식기의 구현과 충분한 학습자료가 요구된다.

본 연구에서는 인식기의 구성을 위해 VQ와 HMM를 사용하며, 인식단위는 음소를 기반으로 하므로 인식 대상어휘인 숫자음 학습자료를 총 16 개의 음소별로 수동분할하여 수집하였다. 인식 숫자음의

음소표기는 표 1과 같다.

표 1: 숫자음과 분할단위

숫자음	기호열
영	yv(여) + vng(엉)
일	i-1(이) + il(일)
이	i-2(이)
삼	s(ㅅ) + a(아) + m (ㅁ)
사	s(ㅅ) + a(아)
오	o(오)
육	yu(유) + uK(욱)
칠	ch(ㅊ) + i-1(이) + il(일)
팔	p(ㅍ) + al(알)
구	k(ㄱ) + u(우)

인식 및 학습 시스템은 크게 세가지 부분으로 구성되어 있다. 첫번째의 전처리 및 음향분석단계에서는 음성신호를 컴퓨터에서 처리하기 위해서 A/D변환을 수행하고, 음성신호를 강조하기 위해서 pre-emphasis단계를 거쳐 음향분석을 수행한다. 두번째 특징 추출단계에서는 음향분석결과를 이용하여 음성인식에 필요한 특징 파라미터를 추출한다. 본 연구에서는 인간의 귀의 특성을 반영하는 주파수 파라미터인 mel-cepstrum계수를 사용하였다.

세번째 단계의 학습시에는 발성숫자음에 가장 근사한 코드북 색인과 모델 종속 코드북에 대한 색인 및 코드워드 색인을 구하고, 이를 사용하여 숫자음 각각에 대한 코드북 색인열을 HMM을 통해 학습하며, 아울러 HMM 각 상태에서의 코드워드 관측확률을 추정하게 된다. 학습단계와는 달리 실험단계에서는 숫자별 HMM모델의 파라미터와 상태별 모든 코드워드의 관측확률을 이용하여 각 상태에서 입력과 가장 근사한 코드워드를 구하고, 이들과 모델에 종속적인 코드워드간의 유사도를 바탕으로 선형결합으로 통해서 모델의 출력을 계산한다. 인식부를 구성하는 모델의 동작원리와 구조에 대해서는 다음절인 3절에서 상세히 기술하고자 한다. 이러한 처리과정을 거쳐서 각 모델에 대한 입력 음성의 출력확률이 구해지면, 이들 값 가운데에서 가장 큰 것을 선택하여 해당모델에 대응되는 단어를 인식결과로 출력한 후 인식을 종료하게 된다. 제안된 시스템의 전반적인 구조와 동작원리는 그림 1과 같다.

3 코드북간의 확률전이 모델

3.1 확률전이 모델과 출력확률 추정

확률전이 모델과 모델내에서의 코드워드의 출력확률 추정을 위해서 학습과정에서 수행되는 처리과정은 다음과 같다. 먼저, 발성숫자음 각각에 대해서 프레임별로 가장 적은 오류를 갖는 코드북 색인 CB_t과 코드워드 색인 CW_t을 아래 식(1)-(4)에 의해서 얻게된다. 이때, 코드북의 갯수를 M, 코드북 i에서

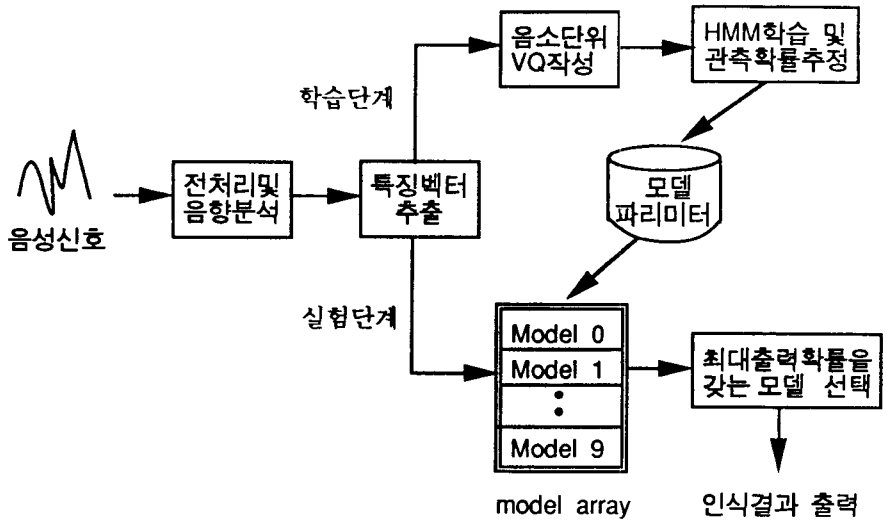

그림 1: 인식 시스템의 동작 및 구조

의 코드워드의 갯수를 N_i, 입력벡터의 길이를 T라고 가정한다. 코드북의 색인열을 얻어낸 후, 인식대상의 모델 m에 종속적인 코드북들만을 사용하여 별도로 각 프레임에 대응되는 가장 적은 오류를 갖는 코드북 색인 CB_t^m 과 코드워드 색인 CW_t^m 을 동시에 얻는다. 이러한 이유는 모델이 나타내는 음소들을 보다 잘 나타내도록 특성화시키기 위해서 이러한 단계가 요구된다. 여기서, D(A,B)는 파라미터 A와 B간의 거리를 나타내는 함수로, euclid 거리를 사용하였다.

$$C_{it} = \arg\min_j D(C_i^j, X_t) \quad (1)$$

$$D_{it} = \min_j D(C_i^j, X_t) \quad (2)$$

$$CB_t = \arg\min_j \{d_{jt}\} \quad (3)$$

$$CW_t = C_{CB_t t} \quad (4)$$

$$D_t = \min_j \{d_{jt}\} \quad (5)$$

$$CB_t^m = \arg\min_l \{d_{lt}\} \quad (6)$$

$$CW_t^m = C_{CB_t^m t} \quad (7)$$

$$where, 1 \le i \le M, 1 \le t \le T, 1 \le j \le N_i, 1 \le l \le \ num\ of\ codebooks\ in\ a\ model\ m$$

결과적으로 전체 코드북과 단어에 종속적인 코드북을 사용한 코드북 색인열을 모두 얻으므로써, 학습자료의 양이 2배만큼 증가하게 됨으로써 보다 모델의 신뢰도를 증가시킬 수 있게 된다. 학습자료의 수집이 완료되면, 인식단어별로 임의의 갯수의 상태를 갖는 HMM 을 학습하게 된다. 본 실험에서는 실

험적으로 상태수를 2로 사용하였다[그림 2].

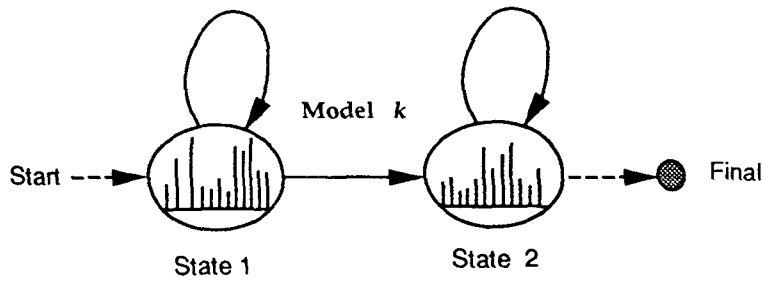

그림 2: 임의의 단어모델 k의 구조

학습이 완료되면 얻어진 P개의 코드북열을 입력으로하여 Viterbi방법을 사용하여 각 상태로 정렬된 코드워드 색인열을 식(8)에 의해서 구한다. 상태별로 정렬된 코드북의 색인을 이용하여 식(9)에 의해서 상태에 의존적인 코드워드분포를 추정하게 된다.여기서, $P_s(w_{ij})$ 는 코드북 i에서 코드워드 j의 관측확률로, HMM에서 얻어진 $P_s(C_i)$와 각 코드북에서 정규화된 코드워드존재확률인 $\bar{P}_{si}(w_{ij})$의 곱으로 계산된다.

$$C_p = \{CB_t\}, 1 \leq p \leq P \quad (8)$$

$$S_p = Viterbi\_Search(C_p) \quad (9)$$

$$P_s(w_{ij}) = P_s(C_i) \times \bar{P}_{si}(w_{ij}) \quad (10)$$

$$where,\ w_{ij} = j-th\ codeword\ in\ a\ codebook\ i$$

실험단계에서는 학습단계에서 얻어진 숫자별 HMM모델의 파라미터와 HMM모델의 각 상태에 대응되는 모든 코드워드의 관측확률을 이용하여 인식을 수행한다. 인식단계는 기존의 VQ+HMM에서 코드워드 색인이 입력되는 것과는 달리 프레임에 대응되는 특징벡터 자체가 입력으로 들어가며, 각 상태에서 가장 가까운 코드워드가 선택된다. 최적의 코드워드만을 고려 하여 계산하는 경우 사용된 코드워드가 해당 모델에 종속적인 코드워드와 연관이 적을 경우 존재확률이 낮으므로, 이러한 점을 보안하기 위해서 모델종속적인 코드워드와의 인접관계를 이용하여 입력벡터의 출력확률을 결정하게 된다[그림 3].

실험단계에서 사용되는 출력식(11)-(13)과 같다. 먼저, t 시점에서 입력과 최적한 코드워드 CW_t를 결정한다. 결정한 코드워드 C_{wt}와 입력벡터간의 거리정보를 이용하여 모델 종속적인 코드워드의 상태 존재확률을 구하게 된다. 여기서, K는 모델 종속적인 코드북의 갯수에 해당되며, $f(x)$는 각 상태에서 최소 거리를 갖는 코드워드의 색인을 의미한다.

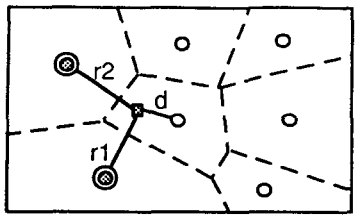

그림 3: state j 에서 입력패턴과 대표패턴간의 관계(2차원의 경우)

$$P_s(X_t) = \frac{1}{K+1} \times [P_s(CW_t) + \sum_{k}^{K} \frac{D(CW_t, X_t)}{D(f(k), X_t)} \times P_s(w_{f(k)k})] \quad (11)$$

$$f(k) = \arg\min_{j} D(C_k^j, X_t) \quad (12)$$

(13)

출력확률을 구한 후, Viterbi 탐색을 통해 모델의 출력을 구하게 된다. 각 모델에 대해서 이러한 연산과정을 거친 후, 가장 큰 출력을 갖는 모델을 최종 인식 단어로 선정하여 출력하게 된다.

3.2 제안된 시스템의 다른 관점

제안된 시스템은 크게 두 가지 관점에서 살펴볼 수 있다. 첫번째로 기존의 FSM모델의 확률적인 형태로 볼 수 있으며, 다른 하나로는 HMM구조를 갖는 인식 모델로 볼 수 있다.

첫번째 관점에서 "확률적"의 의미는 상태간의 천이에만 해당되는 것으로, 각 전이에서 관측심벌은 해당 전이에서의 코드북과 오류가 가장 적은 코드워드로 결정된다. 앞서 설명했듯이, 상태의 전이가 단순히 프레임단위의 각 전이 코드북간의 오류만으로 정의될때 입력에 해당되는 코드워드의 선택이 부적절한 상태전이에서 이루어질 수 있으므로, 이러한 오류를 줄이기 위해서 상태천이를 HMM형태에 학습시켜 전이가 확률적으로 발생되도록 구성된 시스템으로 볼 수 있다.

두번째 관점에서 보면, HMM학습에 있어서 전이확률의 학습과 각 상태에서의 심벌관측확률을 분리하여 학습한다고 볼 수 있다. HMM은 최대 우도 추정(maximum likelihood estimation ; MLE) 학습을 통해서 전이확률과 심벌 관측확률이 동시에 최적화되지만, 제안된 모델의 경우 단순히 코드북의 색인열만을 학습하고 각 상태에서의 심벌관측확률을 단순히 Viterbi방법을 사용하여 정열한 각 상태에서의 코드북의 코드워드의 존재확률로써 사용하므로, 보다 간단한 학습과정이 이루어지게 된다. 이와 아울러 각 상태에서는 최적코드북과 인식 모델에 종속적인 코드북의 최적 코드워드간의 관계를 이용하여 입력 특징벡터의 해당 상태에서의 존재확률을 구한다. 이러한 방법은 최적 코드워드 의 존재확률은 낮으나, 모델 종속적인 코드워드와의 거리가 짧은 경우 모델 종속적인 코드워드의 존재확률을 반영하므로

써 보다 높은 존재확률을 가질 수 있도록 함이다. 이러한 과정을 통해서 단어나 음절과 같은 단위를 음소와 같은 적은 단위로 모델링하는 경우 별도의 음소모델을 작성하기 보다는 VQ와 같은 간단한 과정을 통해 음소에 종속적인 코드북을 작성하고, 상태에 따른 코드북과 코드워드의 존재확률을 사용하여 계산하므로써, 모델 구성이 용이하다고 판단된다.

4 실험 및 결과

본 연구에서 제안된 시스템의 성능평가를 위해서 독립발성된 10개의 숫자음을 대상으로 인식실험을 수행하였다. 40명의 화자(20명 남성/20명 여성)가 4회씩 발성한 숫자음을 대상으로 남여 각각 1회씩의 발성분을 학습에 사용하였으며, 나머지 2회분을 시스템의 성능평가를 위해서 사용하였다. 인식을 위한 음향처리와 특징벡터는 표 2와 같다.

표 2: 실험 조건

item	characteristics
Sample Rate	16 KHz
Pre-emphasis	$1.0\text{-}0.97Z^{-1}$
quantization	16 bits/sample
Feature vector	12 order Mel-scale cepstrum (bilinear transformation)

분할단위별로 독립된 VQ 코드북을 작성하였으며 최적화된 코드북의 작성을 위해 2 가지 초기방법을 사용하여 실험을 수행하였다. 첫번째 방법으로, LBG로 초기화한 후, K-means을 수행하여 코드북을 얻는 방법이 있으며, 두번째로 LBG를 수행한 후 바로 K-means를 사용하여 코드북의 단계가 증가함에 따라 이러한 2단계의 과정을 반복수행한다. HMM의 구현을 위해 128 코드워드를 가진 코드북작성시 첫번째 방법의 경우 오류가 0.5521이였으며, 두번째 경우 0.5518로 수렴된 상태에서의 오류차이는 거의 없었으며, 실제 인식을 수행하였을 경우 1개의 숫자만을 두번째 방법이 더 인식하는 차이만이 존재하였다. 따라서, 이러한 실험결과를 바탕으로 두번째 방법을 사용하여 각 분할단위에 대한 개별코드북을 작성하였다. 분할단위 각각에 대한 학습자료수와 코드워드수 및 코드북의 오류는 표 3과 같다.

다음으로 제안된 시스템의 성능평가에 대해서 기술한다. 제안된 시스템의 성능평가를 위해서 HMM과 FSM형태의 시스템을 별도로 구현하여 성능비교를 수행하였다. 인식결과는 표 4와 같다. 전체 인식율면에서 제안된 시스템이 97.4%, HMM이 94%, FSM이 79%을 보였다. FSM 방식보다는 18.4%의 성능향상을 보였는데, 이는 단순히 frame단위로 출력확률의 크기로 전이하기보다는 확률적인 천이가 보다 효율적임을 알 수 있으므로, 동일한 확률적인 천이를 갖는 HMM에 비해서도 3.4%의 인식율 향상이 있었다. 이는 각 상태에서의 보다 상세한 코드워드 출력확률 모델링이 유효함을 보여준다고 판단된다.

표 3: codebook별 오류와 대표패턴 수

분할기호	학습패턴수	codeword수	오류율
yv	1540	8	0.442
vng	2222	8	0.433
i-1	2073	8	0.414
i-2	4060	8	0.349
il	4627	8	0.379
s	1709	4	0.405
a	3715	8	0.432
m	1959	8	0.387
o	3655	8	0.412
yu	1173	8	0.412
uK	1407	8	0.425
ch	1072	4	0.403
p	735	4	0.392
al	1645	8	0.439
k	699	4	0.398
u	2753	8	0.390

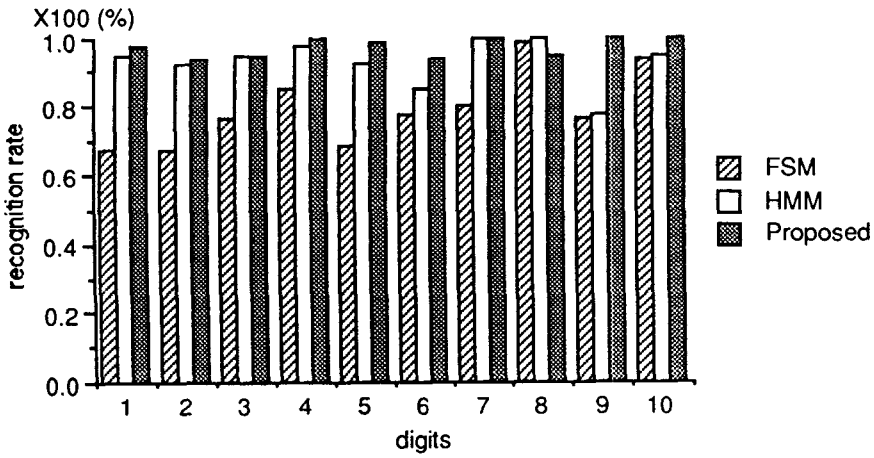

표 4: 제안된 시스템과 FSM 및 HMM의 숫자별 인식율

5 결론 및 검토

본 연구에서는 개별적으로 작성된 음속단위 코드북을 이용하여 코드북 색인의 열을 학습하는 HMM을 이용하여 해당 숫자음을 인식하는 인식모델을 제안하였다. 화자독립 발성된 숫자음을 대상으로한 인식실험 결과, FSM구조의 인식기와 HMM에 비해서 높은 인식율을 나타내는 것을 알 수 있다. 제안된 모델은 부분적으로는 확률적 천이를 갖는 단어모델이며, 상태에서의 출력확률은 학습된 코드북의 색인열로 부터 추정된 출력분포를 이용하며 아울러 모델 종속적인 코드워드를 함께 사용하여 모델의 출력확률을 높이므로써, 인식단어에 대한 모델링 능력을 향상시키므로써 시스템의 성능향상이 있었다고 판단된다.

앞으로의 연구방향으로는 본 논문에서 제안한 모델을 연속숫자음을 대상으로한 인식실험에 적용하여 모델의 유효성을 보이는 연구가 진행되어 할 것이며, 아울러 보다 신뢰성 있는 모델 구현을 위한 제반 이론 연구가 필요하다 판단된다.

참고 문헌

[1] A. Waibel, K. F. Lee, *Readings in Speech Recognition*, Morgan Kaufmann Publisher, Inc., 1990

[2] L. R. Rabiner, B. H. Juang, "An Introduction to Hidden Markov Model", *IEEE ASSP Magazine*, No. 4,pp. 4 - 16, 1986

[3] X. D. Huang, M. A. Jack, "Semi-continuous hidden Markov models for speech signals", *Computer speech and language*, No. 3, pp. 239 - 251, 1989

[4] L. R. Rabiner, B. H. Juang, S. E. Levison, M. M. Sondhi, "Some Properties of Continuous Hidden Markov Model Representations", *AT&T Technical Journal*, Vol. 64, No. 6, pp. 1251 - 1269, July-August 1986

[5] Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantization Design", *IEEE transaction on Communication*, Vol. Com-28, No. 1, January 1980

[6] , D. K. Burton, J. E. Shore, J. T. Buck, "Isolated-Word Speech Recognition Using Multisection Vector Quantization Codebooks", *IEEE trans. on ASSP*, ASSP-33, No. 4, pp. 837 -849, August 1985

[7] S. Huang, R. M. Gray, "Spellmode Recognition Based on Vector Quantization", *Speech Communication*, No.7, pp. 41 - 53, 1988

[8] G. E. Kopec, M. A. Bush, "Network-Based Isolated Digit Recognition Using Vector Quantization", *IEEE trans. on ASSP*, Vol. ASSP-33, No. 4, August 1985