1993년도 한국음향학회 학술논문발표회 논문집(제 12권 1(s)호)

상관성있는 VQ-HMM을 이용한 고립 단어 인식

°이 진수*, 정 광우*, 홍 광석**, 박 병철*
*성균관 대학교 전자 공학과, **제주 대학교 정보 공학과

Isolated Words Recognition using Correlation VQ-HMM

°JinSoo Lee*, KwangWoo Chung*, KwangSeok Hong**, ByungChul Park*
*Dept. of Electronic Engr., Sung Kyun Kwan Univ, **Dept. of Information Engr., Che Ju National Univ.

Abstract

In this paper, we propose the modified VQ,applied correlation between codewords in order to reduce the error rate due to personal and speakers' temporal variation. Such a modified VQ is used in the stage of preprocessing of HMM and the temporal variation is absorbed by nonlinear Decimation and Interpolation of vowel part that we obtain higher recognition rate than not so case. The objects of experiment are Korea 142 DDD regional names and we show that the proposed method increase the recognition rate.

I. 서 론

음성 인식이란 파형 형태의 음성 신호로부터 그 속에 포함된 정보를 컴퓨터에 의해서 자동적으로 추출해내는 작업을 말한다. 그러나 음성의 특징은 사람마다 다르며 같은 사람이 같은 말을 하여도 발성 속도나 길이 등이 다르고 ,또한 연속해서 발성할 경우에는 음운 경계 사이에서 조음 결합 현상이 발생함으로 연속음이나 화자 독립 인식시에는 어려움이 따른다.

현재의 음성 인식 기술로 많이 쓰이고 있는 방법에는 DTW, VQ, 신경 회로망, 퍼지, HMM 등이 있다. 이러한 방법들중에서 DTW는 높은 인식률을 얻을 수 있다는 장점이 있으나 많은 수의 기준 패턴을 구성하기 위해서는 많은 시간과 노력이 필요하고 인식 과정에 걸리는 시간이 길기 때문에 대용량 단어의 인식인 경우에는 실시간 처리가 어렵고 하드웨어 구성이 어렵다는 문제점이 있다. 또, VQ는 계산량이 적고 인식률이 높다는 장점이 있으나 벡터 양자화에 의한 효과로 인하여 대이타가 국부적으로 손상될 수 있다. 최근에는 신경회로망이나 퍼지이론을 적용하는 방법이 상당히 관심을 끌고 있다. 신경회로망은 자체적인 학습기능을 가지고 있어 훈련 패턴에 비해 약간 다른 패턴이 들어 올지라도 잘 인식할 수 있고 병렬적인 구조를 가지고 있으므로 처리 속도가 빠른 장점을 가지고 있다. 그러나 훈련시에 상당히 많은 시간이 필요하고 적용에 따라 메모리가 많이 사용되고 하드웨어 지원이 어렵다는 단점이 있다.

이에 반해 HMM은 기본 단위로서 단어 이하의 음성 인식 단위, 예를 들어 음절이나 음소, diphone 등을 사용하기 때문에 단어나 문장등의 모델을 쉽게 구성할 수 있다. 이 방법은 인식 소요 시간이 짧고 음성 신호의 국부적인 변화를 잘 나타낼수 있을 뿐 만 아니라 통계적인 변수를 사용하기 때문에 비교적 적은 기억 용량으로도 원하는 음성 신호의 각 패턴들을 기억시킬 수 있다. 또,HMM 은 다른 모델에 비해 안정되고 간결하게 음성을 모델링하는 것이 가능하다. 최근 음성 인식의 연구 방향은 HMM을 이용한 대어휘 고립 단어 인식, 연속 숫자음 인식, 연속 문장 음성 인식으로 연구의 방향을 확장해가고 있다.어서의 음성 인식의 방향은 연속 숫자음 인식, 연속 문장 음성 인식으로 연구의 방향을 확장해 가고 있다.

그러나 기존의 HMM은 시간적 변동의 차이로 인해 인식률이 저하될 수 있고 이를 방지하기위해 하나의 인식 단위에 대해서도 여러개의 모델을 사용한 경우도 있으나 이는 많은 계산 시간이 요구되는 단점이 있다. 본고에서는 이를 해결하고자 이산 HMM 전처리 단계인 벡터 양자화 단계에서 상관성있는 코드북을 작성하여 시간적인 길이의 문제점을 해결하기위하여 변형된 VQ-HMM을 이용한 고립 단어 인식에 관한 방법을 제안한다. 즉, 코드북 사이에 상관성을 적용함으로서 음성의 안정된 부분의 벡터 양자화 레벨도 역시 안정되게 출력되도록 하였다. 이렇게 함으로서 안정된 부분,특히 모음 부분에서 비선형적인 데시메이션과 인터폴레이션을 통해 시간 변동을 흡수함으로서 인식률을 향상시키는 방법에 대하여 기술한다. 본 실험에서는 기존의 HMM 방법과 변형된 VQ-HMM을 이용한 인식률 비교를 통하여 제안된 방법의 타당성을 검증한다.

II. 분석 파라미터

켑스트럼은 파형의 소구간 스펙트럼 X(w) 의 진폭에 대수를 취하고 이를 역 FFT 한 것으로 정의되고 성도 전달함수와 여기 성분을 분리시켜 얻을 수 있는 특징이 있는 것으로 알려져 있다.

다음 그림 2-1 을 음성 생성의 선형 분리 등가 모델로 생각한다면 출력 음성 x(t)는 유사 주기적인 입력 파형 g(t) 에 의해 성도 필터가 구동된 응답이라고 생각할수가 있으며, 성도의 임펄스 응답을 h(t) 라고 한다면 convolution 정리에 의해 다음과 같은 식으로 나타낼수 있다.

$$x(t) = \int_0^t g(\tau)h(t-\tau)d\tau \qquad (2\text{-}1\text{-}a)$$

$$X(\omega) = G(\omega)H(\omega) \qquad (2\text{-}1\text{-}b)$$

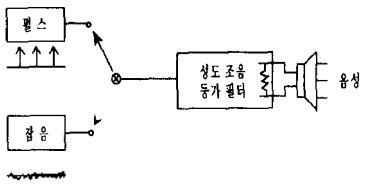

그림 2-1. 음성 생성의 선형 분리 등가 모델

식 2-1-b 의 진폭에 대수를 취하면

$$\log|X(\omega)| = \log|G(\omega)| + \log|H(\omega)| \qquad (2\text{-}2)$$

식 2-2 의 식을 역 FFT 취한것이 FFT 캡스트럼 계수 c(n) 이다. 즉,

$$c(n) = F^{-1}\log|X(\omega)| = F^{-1}\log|G(\omega)| + F^{-1}\log|H(\omega)| \qquad (2\text{-}3)$$

캡스트럼 분석의 특수한 형태로서 선형 예측 분석법에 의해 추정된 전달함수

$$H(z) = \frac{1}{1+\sum_{i=0}^{p}\alpha_i Z^{-i}}$$ 를 필터의 전달 함수로 하여

계수를 추출할 수 있다.

이 계수는 선형 예측 모델에 의해 추출되었으므로 LPC 캡스트럼 계수라고 부르고 LPC 계수 α_n로부터 직접 추출할 수가 있으므로 FFT 과정이 불필요하여 추출 시간이 빠르다는 장점이 있다. LPC 계수로 부터 캡스트럼 계수는 다음식에 의해 순환적으로 계산될 수 있다.

$$c(n) = \alpha_n + \sum_{k=1}^{n-1}\left(\frac{k}{n}\right)c(k)\alpha_{n-k} \qquad (2\text{-}4)$$

III. 상관성을 고려한 VQ

벡터 양자화는 효과적인 데이타 압축 기법으로서 이산 HMM 의 전처리 단계로서 사용된다. VQ 를 수행할 때 가장 중요한 것은 입력 벡터와 해당하는 코드워드와의 왜곡이 가장 적도록 효과적으로 코드 북을 작성하는 것이다. 코드북 작성 알고리즘으로는 k-means 알고리즘,LBG 알고리즘, mixture density estimation알고리즘등이 있다. 그러나 기존의 VQ 기법에는 코드북사이의 상관성을 고려하지 않았기 때문에 거의 안정된 벡터들의 코드가 불안정하게 출력되어 인식률을 저하시킬수 있다. 이에대해 본 논문에서는 기존의 방법처럼 랜덤하게 설정하는것이 아니라 코드 북 레벨 사이에 상관성을 고려하여 인접 코드북은 상관성이 크게 되도록 작성하는 방법을 사용한다. 상관성 측정으로 캡스트럼 거리를 사용하였다. 이러한 방법으로 코드북을 작성하면 중간의 안정한 부분에서의 벡터 양자화 계수는 그렇지 않은 방법보다 안정되게 나타난다.

일반적으로 특징 벡터 계수들 사이의 거리 계산은 다음과 같이 정의되는 유클리디안 거리를 사용한다.

$$d_{cep} = \sqrt{\sum_{i=1}^{p}\left(C_t(i)-C_r(i)\right)^2} \qquad (3\text{-}1)$$

여기서 $C_t(i)$ 와 $C_r(i)$ 는 계산되어질 캡스트럼 계수이고 P 는 차수이다.

그러나 본 논문에서는 보다 좋은 인식률을 얻기위해서 계수의 차수마다 중요성에 해당하는 가중 함수를 사용한다. 가중치는 각 차수의 계수의 분산을 구하여 그분산의 역수를 취하여 준다. 분산은 계수값들의 분포도를 나타내는데 분산이 작다는 것은 계수값이 일정한 범위에 집중되어 있다는 것을 나타내므로 그 중요도가 크다고 볼수 있다. 따라서 이에 따른 거리는

식 3-2 와 같다.

$$d_{cep} = \sqrt{\sum_{i=1}^{p} W(i)\left(C_t(i)-C_r(k)\right)^2} \qquad (3\text{-}2)$$

여기서 $W(i)$ 는 각 차수의 가중 함수이다. 다음 그림은 각 차수의 분산도를 나타내는 그림이다. 그림 3-1 에서 알수 있듯이 1차와 10차의 계수가 큼을 알수 있다.

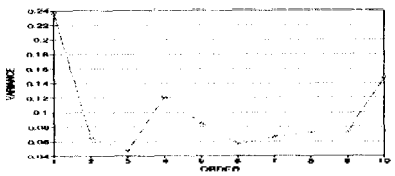

그림 3-1. 각 차수의 분산도

이 거리값을 이용하여 원점으로부터 거리값이 가장 큰 벡터를 레벨 0 에 놓고 이 벡터와 가장 거리차가 적은 벡터를 레벨1 에, 또 레벨 1 과 거리차가 가장 적은 벡터를 레벨 2 에, 등등으로 놓는 방법으로 작성하였다. 다음 그림은 상관성을 고려하지 않았을때와 고려했을 경우 모음 /이/ 에 대한 벡터 양자화 계수를 나타냈다. 그림에서 알수 있듯이 안정된 모음 구간에는 상관성을 고려한 것이 역시 안정됨을 알수있다.

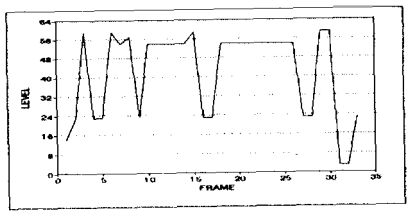

그림 3-2. 상관성을 고려하지 않았을 경우 /이/

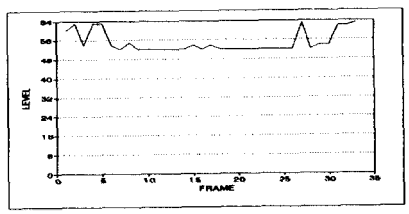

그림 3-3. 상관성을 고려하였을 경우 /이/

IV. 이산 HMM

HMM 은 관측이 불가능한 처리를 관측이 가능한 심벌을 발생시키는 다른 처리를 통하여 추정하는 이중의 확률처리이다. 즉, 주어진 관측열을 O(1),O(2),- - O(t)이라하면 이는 임의의 음성 모델에서 발생되어졌다고 생각되어질수 있다. 성도(Vocal track)가 유한한 몇개의 상태로 구성되어 있고 각 상태에서는 그 상태에 연관된 기본적인 소신호 신호를 발생시키고 이 소신호 신호들의 시간적인 변화는 Markov Chain 의 상태 천이 확률 밀도에 의하여 결정되어 질수 있다. 경험적으로 볼때 음성신호는 Markov Process적인 특성을 가지고 있으므로 위에서 설명한 모델을 만드는데 적합한것으로 생각되어 진다.

HMM을 모델화하는데 있어서는 다음의 세 가지의 확률 parameter 를 결정하여야 한다.

$$\lambda = (A, B, \Pi)$$

여기서 λ는 임의의 모델, A 는 어떠한 상태에서 다른 상태로 이동할 천이 확률이고 B는 어떠한 상태에서 관측열이 존재할 확률, Π는 관측열이 임의의 상태에서 시작할 확률이다.

임의의 관측열과 모델이 주어졌을 때 이 모델에 의해서 주어진 관측열이 발생될수 있는 확률을 구하는 문제는 forward-backward 알고리즘으로 해결할 수 있고 가장 잘 설명할 수 있는 경로를 찾는 문제는 Viterbi 알고리즘으로 찾을 수 있다.

HMM 에서 가장 중요한 문제는 모델 파라미터를 최적하게 훈련시키는 문제로서 아직 해석적인 알고리즘은 없지만 Baum-Welch 의 반복적인 알고리즘이 일반적으로 이용된다.

시간 t 에서 상태 i 에 있고 시간 t+1 에서 상태 j 에 있을 확률을 $\gamma_t(i,j)$라 하고 시간 t 에서 상태 i 에 있을 확률을 $\gamma_t(i)$라고 하면 수행 과정은 다음과 같다.

1 단계 : $\pi_i = \gamma_1(i)$

2 단계 : $a_{ij} = \dfrac{\sum_{t=1}^{T-1} \gamma_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$

3 단계 : $b_j(k) = \dfrac{\sum_{t \in O_t = v_k} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)}$

다음 그림 4-1은 전체적인 VQ-HMM 인식기의 블록 다이어그램이다.

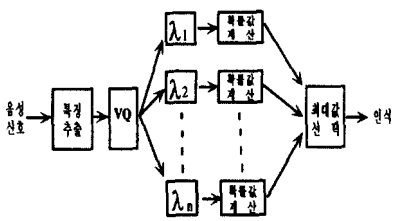

그림 4-1. VQ-HMM 인식도

입력된 음성 신호의 동적 영역을 늘이기 위해서 Preemphasis 를 행하고 음성을 소구간 처리를 하기위해 window 를 취하여 몇개의 블럭 (프레임)으로 나누고, 각 프레임마다 음성의 특징 파라미터인 LPC 계열 계수, 여기서는 LPC- 켑스트럼을 추출한다. 벡터 양자화부는 입력 벡터와 코드워드와의 거리차를 계산하여 거리차가 가장 작은 코드북의 레이블을 출력시켜 관측열을 생성시킨다. 확률값 계산부는 이미 트레이닝된 각각의 HMM 모델이 전단의 벡터 양자화부에서 생성된 관측열이 생성될 확률을 계산한다. 확률값 계산은 forward 알고리즘으로 수행하고 최대값 선택부에서 계산된 확률값들 중에서 최대인 값을 취하여 인식 음성으로 판단한다.

상관성있는 벡터 양자화기를 사용함으로서 입력된 벡터열과 이미 훈련된 벡터열과의 시간 변동은 다음의 방법을 사용함으로서 흡수될 수 있다.

음성의 길이는 전적으로 자음보다는 모음의 길이에 좌우되고 모음의 특징 파라미터는 거의 안정되게 나온다. 시간 변동은 이 모음부분에 비선형적인 처리를 함으로서 해결된다. 즉, 입력 벡터열이 훈련 벡터열의 길이보다 길때는 중간 모음부분에서 데시메이션을 취하고 짧을 때는 인터폴레이션을 취한다. 음절의 중간 1/3 부분을 모음이라고 가정하고 실험을 하였다.

V. 실험 및 고찰

음성 데이타 환경	10 kHz 샘플링, 8 bit 양자화 486 PC, 잡음 환경
분석 프레임	10 msec, 겹침없이
기본 파라미터	LPC-CEPSTRUM 10 차
벡터 양자화 레벨	64 레벨
모델의 상태수	5 개

표 5-1. 실험 환경

실험 환경은 표 5-1 에 나타나있고 사용된 모델은 그림 5-1 에서 나타난 것과 같이 상태수 5 을 가진 단어 단위 LEFT TO RIGHT 모델을 사용하였다.

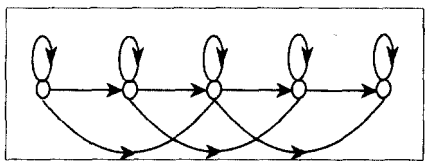

그림 5-1. LEFT TO RIGHT 모델

한국어 DDD 지역명 142 개를 남성 화자 1 인이 5 번씩 발성한 것을 대상으로 해서 기존의 방법과 본 논문에서 제안한 방법을 비교해서 실험해보았다. 실험 결과로 제안한 방법의 인식률은 100%로서 그렇지 않은 방법보다 8.7%의 인식률의 증진을 보였다. 기존의 방법에서 지역명의 모음부분이 같아도 발성 길이에 따라 잘못 인식되는 경우가 있었다. 예를 들어 /청주/는 /정주/로 인식했으며 /순천/은 /춘천/등으로 인식하였다. 즉, 인식률은 모음에 크게 의존하는 것을 알수 있

었다.

이 결과를 통해 제안한 방법이 모음의 안정 부분을 충분히 흡수해서 인식률을 증진시킴을 알수 있다.

서울	부산	대구	인천	광주
대전	수원	가평	강화	구리
김포	문산	발안	성남	안성
안양	양평	여주	용인	원당
이천	천곡	평택	포천	춘천
강릉	동해	삼척	속초	양구
영월	원주	인제	정선	철원
태백	평창	홍천	화천	횡성
청주	괴산	단양	보은	영동
옥천	음성	제천	진천	충주
공주	금산	논산	당진	대천
부여	서산	서천	예산	온양
천안	청양	홍성	경주	고령
구미	군위	김천	봉화	상주
성주	안동	영덕	영양	영주
영천	예천	왜관	울릉	울진
의성	점촌	청도	청성	포항
하양	현풍	마산	거제	거창
고성	김해	남해	밀양	사천
산청	양산	울산	의령	진주
진해	창녕	충무	하동	함안
함양	합천	전주	고창	군산
김제	남원	무주	부안	순창
이리	임실	장수	정주	진안
강진	고흥	곡성	광양	구례
나주	담양	목포	무안	보성
순천	여수	영광	영암	완도
장성	장흥	진도	함평	해남
화순	제주			

표 2. 음성 데이타

VI. 결 론

개인의 환경에 따른 시간 변동을 흡수하는 방법으로서 코드 워드사이에 상관성을 고려한 방법을 제안, 실험하였다. 이러한 양자화기를 사용하고 모음 부분에서 시간차나는 만큼의 데시메이션과 인터폴레이션을 통해 시간 길이를 맞춤으로서 인식률이 8.7 % 증진되었다.

참고 문헌

[1] L.R.Rabiner, B.H.Juang, "An introduction to Hidden Markov Models", IEEE ASSP MAGAZINE, January 1986

[2] L.R.Rabiner," A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition", PROCEEDING OF THE IEEE, Febr-uary 1989

[3] Robert M.Gray "Vector Quantization", IEEE ASSP MAGAZINE, April 1984

[4] G.David Forney, JR, "The Viterbi Algorithm" , PROCEEDING OF THE IEEE, Vol.61, NO.3, March 1973

for speech processing IEEE on ASSP, vol 24, Oct.1976

[7] 유 현창, 박 병철 "음소 길이를 고려한 3-State Hidden Markov Model 에 의한 한국어 음소 인식 ", 성균관 대학교 석사 학위 논문 , 1988

[8] 이 희정, 박 병철 " 음소 단위 음성 인식을 위한 자동 세그멘테이션과 단어 가설", 성균관 대학교 박사 학위 논문, 1989

[9] X.D.Huang, Y.Ariki, M.A.Jack, "Hidden Markov Models for Speech Recognition", EDINBURGH University Press , 1990

[10] L.R.Rabiner, R.W.Schafer, "Digital Processing of Speech Signals ", Prentice-Hall, 1978

[11] Shuzo Saito, Kazuo Nakata, "Fundamentals of Speech Signal Processing", Tokyo University, 1985