1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

중규모급 단어 인식기의 실시간 구현을 위한 무감독 단어집단화 알고리듬

임동식*, 김진영*, 백성준**
전남대학교 전자공학과*,
고품질 전기·전자 부품 및 시스템 연구센터
서울대학교 전기공학부**

Unsupervised Word Grouping Algorithm for real-time implementation of Medium vocabulary recognition

Dong Sik Lim*, Jin Young Kim*, Seong Joon Baek**
Dept. of Electronics Engineering, Chonnam National Univ* &
Research Center for High-Quality Electric Components and Systems
(E-mail : dslim@dsp.chonnam.ac.kr, kimjin@dsp.chonnam.ac.kr)
Dept. of Electrical Engineering, Seoul National Univ**

요 약

본 논문에서는 중규모급 단어인식기의 실시간 구현을 위한 무감독 단어집단화 알고리듬을 제안한다. 무감독 단어집단화는 인식대상 어휘 수가 많은 대용량 음성인식 시스템에서 대상 어휘 수를 줄여주는 역할을 하는 전처리기의 성격을 갖는다. 무감독 집단화를 위해 각 단어의 유·무성음 고유의 특성을 잘 반영할 수 있는 특징 파라미터 5개를 사용하여 패턴 인식과 회귀분석에서 널리 사용되고 있는 분류·회귀트리(Classification And Regression Tree)에 적용시키는 방법으로 접근하였고, 각 단어의 frame 수를 일정하게 n개로 분할(segment)하여 1개의 tree를 생성시키는 방법과 각 segment에 해당하는 tree를 생성시켜 segment들 사이의 교집합 성분으로 단어들을 집단화 하였다. 실험결과 탐색 대상단어 22개에서 평균2.21개로 줄어 전체 대상 단어의 10%만을 탐색하여 인식할 수 있는 방법을 제시할 수 있었다.

I. 서론

음성 언어 처리를 위해서는 대용량 어휘의 실시간 연속음성인식에 관한 연구가 필수적인데, 고립단어인식은 현재 수십에서 수만 단어 정도의 어휘에 대해서는 신뢰할 수 있는 단계까지 와 있으나 많은 계산량에 따른 어려움이 여전히 존재한다. 이러한 문제점을 해결하고자 지금까지 고립단어 음성인식에 있어서 효율적인 계산량 감축방법의 접근방법으로 HMM을 기반으로 비터비 빔(Viterbi Beam)탐색기법이 사용된 바 있다[1-2]. 과거 국내의 한 보고서에 의하면 대용량 단어 인식 시스템을 위한 전처리기를 제안하였는데 이는 입력된 단어를 4개의 음소군(파·마찰음, 공명음, 정지음, 모음)으로 나누어주는 음소군 분류 시스템이 있는데 이는 태깅(tagging)작업이 선행되어야 하는 번거로움이 있다[3].

본 논문에서는 각각의 단어를 잘 표현해 줄 수 있는 특징벡터(feature vector)를 이용하여 각 단어를 이루는 유성음과 무성음의 특성에 의해 구별되어질 수 있는 고유 파라미터 값에 따라 집단화하는 것을 목적으로 한다.

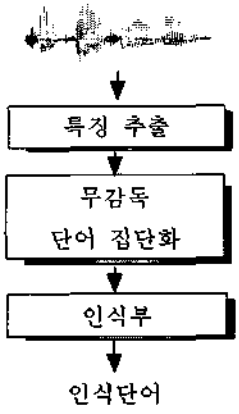

그림1. 무감독 단어 집단화에 의한 인식

II . HMM을 이용한 음성인식

2.1. 연속분포 HMM의 학습

특징추출은 8kHz,16bits로 A/D한 후, $(1-0.97z^{-1})$의 전달함수를 갖는 필터를 사용하여 pre-emphasis된 후 25msec 길이의 프레임(frame)단위로 분할되고 15msec씩 중첩된다. 특징 파라미터는 12차 mel-cepstrum, 1차 normalized log energy, 12차 delta-cepstrum, 1차 delta-energy를 사용하여 각 frame에 26차 파라미터를 사용하였다.

본 논문에서는 연속분포 HMM을 사용하여 고립단어 인식을 수행하였으며 8개의 State, 3개의 Mixture를 갖는 Gaussian 혼합밀도 함수로 출력 확률분포를 추정하였다.
HMM 학습과정은 그림2와 같다.

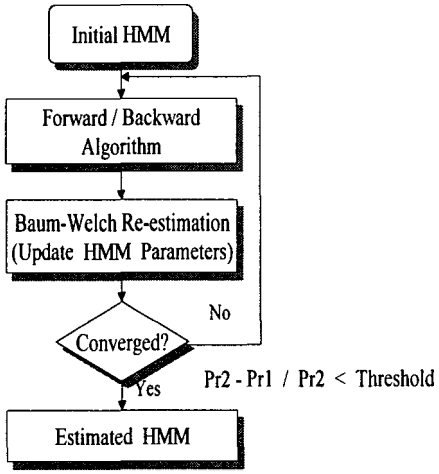

그림2. HMM 학습과정 흐름도

그림2에서 재 추정된 값이 임계값에 수렴하거나 주어진 훈련회수가 될 때까지 반복하다가 최종 HMM 모델 파라미터가 생성된다. 여기서, Pr2는 현재 확률, Pr1은 이전 확률을 뜻한다.

2.2. 학습을 통한 인식

본 논문에서 학습모델을 바탕으로 한 인식실험은 forward 알고리듬으로 수행하였다. 자동차 항법 장치 시스템을 위한 고립단어 22개에 대하여 52명 화자의 발성 단어를 가지고 학습을 한 후, 실험은 18명 화자의 발성 단어를 가지고 하였다. 총 396개의 단어에 대해 392개를 정확하게 인식 해 98.98%의 비교적 높은 인식률을 보였다.

또한 파라미터의 중요도를 살펴볼 수 있는 방안으로 delta cepstrum과 delta energy만을 특징 파라미터로 사용하였을 경우 인식률은 94.7%를 나타냈고, mel cepstrum과 log energy만을 특징 파라미터로 사용한 경우 인식률은 96.21%를 보였다. 이상에서 볼 수 있듯이 파라미터 중요도의 관점에서 볼 때 mel cepstrum은 중요한 파라미터로 여길 수 있다.

III. 무감독 단어 집단화 방법

무감독 단어 집단화를 위해서는 몇 가지 조건을 만족해야 한다.

첫째, 단어 집단화에 필요한 계산량이나 파라미터 수는 많지 않아야 한다.
둘째, 집단화를 통해서 추출된 후보 단어의 수는 전체 대상 단어보다 훨씬 적어야 한다.
셋째, 축소된 후보 단어 때문에 원래의 음성 인식 시스템 성능이 떨어져서는 안된다.

본 논문에서 제안하는 무감독 단어 집단화 방법은 이러한 조건을 바탕으로 각 단어를 이루는 유성음과 무성음의 특성에 의해 구별되어 질 수 있는 파라미터 값에 따라 집단화하는 것을 그 목적으로 한다.

3.1. 특징 파라미터 선정 및 추출

단어 집단화를 위한 특징 파라미터는 5개를 사용하였는데, 그 기준은 유·무성음의 특징이 비교적 잘 구별지

어 질 수 있는 고·저주파 대역의 에너지 비율, 영 교차율, filter bank 저주파 대역 및 고주파 정규화 로그 에너지 비율과 각 frame에서의 정규화 로그 에너지이다.

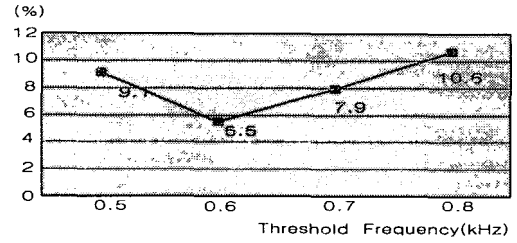

그림3.1 고주파에너지 / 저주파에너지 겹침 비율

고주파 및 저주파 대역의 에너지 비율을 나타내기 위한 준비작업으로 고·저주파를 구분할 수 있는 경계 주파수를 찾아야한다. 그림3.1에서 고주파와 저주파 경계의 임계주파수가 0.6kHz일때 최소의 overlap을 보였다. 이 경우 유성음은 194/3226=6.0% , 무성음의 경우 19/646=2.9%의 겹침 현상을 보였다. 따라서, 고주파와 저주파의 경계주파수는 0.6kHz로 선택하였다. 그림3.2는 "스포츠"에 대한 각 파라미터의 특징추출도를 보인다.

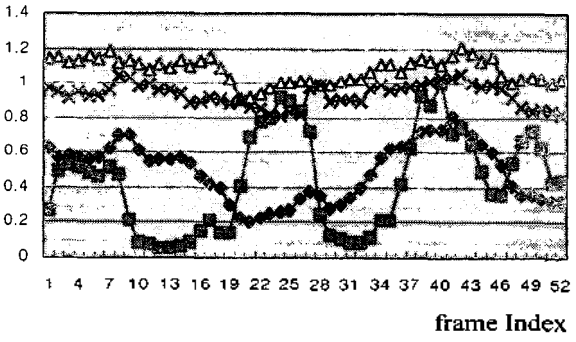

그림3.2 "스포츠"에 대한 특징추출도

△ : filter bank 저주파 대역 정규화 로그에너지 비율

× : filter bank 고주파 대역 정규화 로그에너지 비율

◇ : 영교차율

□ : 정규화 로그에너지

filter bank 저주파 대역의 정규화 로그 에너지 비율

: 1~2.5kHz / 125~750Hz

filter bank 고주파 대역의 정규화 로그 에너지 비율

: 3~4kHz / 1~2.5kHz

실험에 적용된 frame은 25msec의 시간을 차지하는데 각 frame, 즉 25msec의 시간동안 유성음 영역에서의 영교차 회수는 평균55.6회(0.278), 무성음 영역에서의 영교차 회수는 122회(0.61)로 유·무성음이 비교적 명확히 구별되었다. 영교차율은 각 frame의 200샘플에 대한 level crossing 회수를 적용하였다.

3.2. CART방법에 의한 집단화

분류·회귀 트리(classification and regression tree)라 불리우는 CART는 크기의 순서가 정해져 있는 ordered 변수 이외에도 categorical 변수에 대해서도 패턴인식 및 회귀분석을 적절히 수행할 수 있다. CART는 변수가 연속형, 다시 말해서 실 변수(real-valued variable)인 경우이면 $x \leq c$ 형태의 질문을(x:특징변수), 이산형인 경우에는 x가 가능한 모든 종류 부분집합에 속하는 경우를 고려하여 질문을 하게된다[4-5]. 또한, 실 변수의 경우 $x \leq c$패턴의 질문형태에서 하나의 특징변수 x가 아닌 2개 이상의 특징변수의 선형조합(linear combination)도 가능한데 이러한 형태를 CART-LC(Linear Combination)라고 한다. CART-LC는 하나의 노드에서 이루어지는 질문이 선형조합이기 때문에 CART에 비해 일반적으로 더 적은 노드를 생성시키는 경향이 있다.

본 논문에서는 무감독 단어 집단화를 위해 각 단어가 가지는 frame수를 일정하게 n개로 분할(segment)하여 각 segment에 대한 5개의 특징 파라미터의 평균을 구한 후 n×5개의 파라미터를 CART의 입력으로 하여 1개의 tree를 생성하여 집단화를 한 경우와 n개 segment 각각에 대해 tree를 생성시켜 n개 segment들 사이의 교집합 성분으로 단어들을 집단화 하는 2가지의 방법을 실험하였으며, 또한 CART, CART-LC에 적용시켜 하나의 노드가 포함하는 단어의 평균 개수에 의해 집단화의 성능 평가 및 비교를 할 수 있다. 집단화에 사용된 음성데이터에서 학습과 실험은 2회 중복화자를 허용하여 녹화자 52명이 발성한 22개의 단어를 가지고 적용하였다.

또한, 5개 각 파라미터에 대한 평균과 표준편차를 구한 후 Z-Score를 식 1과 같이 적용하였다.

$$\text{Z-Score} : \frac{\chi - \mu}{\sigma} \qquad (1)$$

다차원적인 변수들을 축소, 요약하고 서로 상관되어 있는 반응변수들을 적절히 변환시켜 소수 몇 개의 의미있는 주성분을 유도·해석을 대상으로 하는 방법을 사용하여 Z-Score방법과 비교한다.

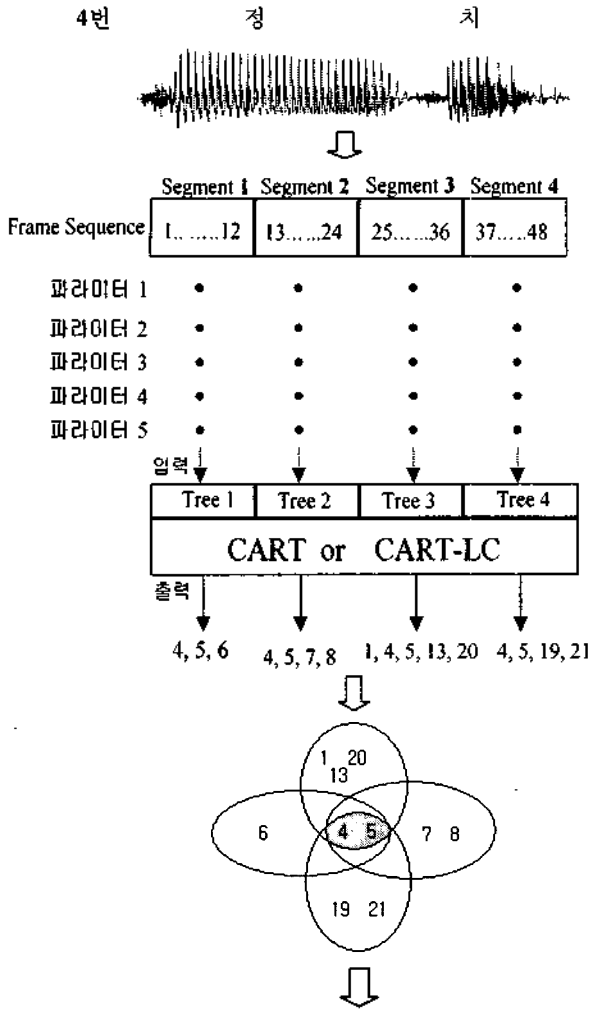

그림3.3 각 segment에 대한 교집합 셋 추출과정

그림3.3은 4번목록의 단어가 입력될 때 4개의 segment에 대한 교집합 셋으로 4,5번이 추출된 경우를 보이고 있다. 이는 연속분포 HMM기반의 인식실험시 4번 단어가 테스트 단어로 입력될 때 전체 22개 단어중 4,5번 단어만을 검색하여 인식결과를 보여주게 된다.

표1. 평균 탐색 단어후보 개수 및 성능비교

적용 \ tree 개수	tree 1개(Node 수)	tree 4개
CART	6개 (69)	2.74개
CART-LC	5.5개 (48)	2.21개

표2에서는 각 파라미터들의 평균과 표준편차를 이용하는 Z-Score방법과 주성분 분석으로 5개 특징파라미터의 96%에 해당하는 중요도를 반영하는 PCA(Principal Component Analysis)를 통해 분석한 결과를 보였다. 각 segment의 파라미터 평균값들에 대해 Z-Score를 적용한 후 CART방법에 의해 단어 집단화할 경우 가장 나은 결과를 보였다.

표2. Z-Score 와 PCA의 적용

	집단화 단어 평균개수	집단화로 인한 탐색 축소율(%)
Z-Score 적용	3.97개	82%
PCA 적용	5.51개	75%
Z-Score 적용 후 PCA 적용	6.45개	71%

IV. 결론

본 논문에서는 중규모급 고립단어 인식시스템을 위해 각 단어의 유·무성음 고유의 특성을 잘 반영할 수 있는 특징 파라미터 5개를 사용하여 무감독 단어 집단화 알고리듬을 소개했다. CART-LC에 의해 4개의 tree를 생성시키고 교집합 셋에 의한 단어들의 집단화 학습결과 전체 대상단어 22개중 평균 탐색단어 후보 수가 2.21개로서 탐색 어휘수를 약90% 줄였으며, 앞으로 무감독 단어집단화를 기반으로 연속분포 HMM을 이용한 1000^{+}단어 실시간 인식시스템을 구현할 예정이다.

참고문헌

[1] X. D. Huang, Y. Ariki, and M. A. Jack, HMM for Speech Recognition, Edinburgh University Press, Edinburgh, England 1990.

[2] N.Y. Han, H.R. Kim, K.W.Hwang, Y.M. Ahn, J.H. Ryoo, A Continuous Speech Recognition System Using Finite State Network and Viterbi Beam Search for the Automatic Interpretation, Proc. ICASSP 1995, Vol. 1, pp.117-120

[3] 한국 과학기술원, 한국어 음성인식 시스템 개발 연구, 제 5차년도 최종보고서,1989

[4] Sreerama K.Murthy, On Growing Better Decision Trees from Data, the degree of Doctor of Philosophy, Baltimore, Maryland 1995.

[5] L. Breiman, J.H. Friedman, R. A. Olshen, and C.J.Stone, Classification and Regression Trees, Wadsworth Statistics/Probability Series, Belmont, CA,1984.