한국음향학회지 제25권 제7호 pp. 319-324 (2006)

가변어휘 핵심어 검출 성능 향상을 위한 비핵심어 모델

Non-Keyword Model for the Improvement of Vocabulary Independent Keyword Spotting System

김 민 제*, 이 정 철*
(Min-Je Kim*, Jung-Chul Lee*)

*울산대학교 컴퓨터 정보통신 공학부

(접수일자: 2006년 9월 7일; 수정일자: 2006년 10월 12일; 채택일자: 2006년 10월 24일)

본 논문에서는 화자독립 가변어휘 핵심어 검출기의 성능을 개선하기 위하여 두 가지의 새로운 비핵심어 모델링 방법을 제안한다. 첫째는 K-means 알고리즘 기반 monophone 군집화 방법을 개선하기 위해 monophone을 state단위로 결정트리를 기반으로 군집화하여 비핵심어를 모델링하는 방법이다. 둘째는 single state multiple mixture 방법을 개선하기 위해 음절단위 multi-state multiple mixture 방법으로 모델링하는 방법이다. 실험에서 ETRI 표준 한국어 공통 음성 단어 DB를 이용하여 트라이폰 모델을 훈련하였고, 훈련에 사용하지 않은 음성데이터를 이용하여 핵심어 검출 closed 테스트를 수행하였다. 그리고 사무실 환경에서 4명의 화자가 각각 100문장씩 발성한 400문장의 음성데이터를 이용하여 100단어 핵심어 검출 open 테스트를 수행하였다. 실험 결과 결정트리기반 상태 군집화 방법이 기존의 K-means 알고리듬 기반 monophone clustering 방법보다 핵심어 검출 성능이 28%/29%(closed/open test) 향상되었다. 그리고 음절단위 multi-state multiple mixture 방법이 비핵심어 전체를 single state 모델로 구성하는 방법보다 핵심어 검출 성능이 22%/2%(closed/open test) 향상됨으로써 본 논문에서 제안한 두 가지 알고리듬이 우수한 결과를 나타내었다.

핵심용어: 핵심어 검출, 비핵심어 모델, 결정트리기반 상태 군집화, HMM

투고분야: 음성처리분야 (2.5)

We propose two new methods for non-keyword modeling to improve the performance of speaker- and vocabulary-independent keyword spotting system. The first method is decision tree clustering of monophone at the state level instead of monophone clustering method based on K-means algorithm. The second method is multi-state multiple mixture modeling at the syllable level rather than single state multiple mixture model for the non-keyword. To evaluate our method, we used the ETRI speech DB for training and keyword spotting test (closed test). We also conduct an open test to spot 100 keywords with 400 sentences uttered by 4 speakers in an office environment. The experimental results showed that the decision tree-based state clustering method improve 28%/29% (closed/open test) than the monophone clustering method based K-means algorithm in keyword spotting. And multi-state non-keyword modeling at the syllable level improve 22%/2% (closed/open test) than single state model for the non-keyword. These results show that two proposed methods achieve the improvement of keyword spotting performance.

Key words: Keyword Spotting, Non-keyword model, Decision Tree-based state clustering, HMM

ASK subject classification: Speech Signal Processing (2.5)

I. 서 론

가변어휘 인식은 인식대상 어휘를 추가 또는 변경하고자 할 경우 미리 정의해둔 subword모델들의 결합으로 인식대상 어휘를 구성할 수 있으므로 DB의 재 수집 및 재훈련이 필요 없는 장점이 있다. 따라서 최근 단어인식에서는 고립단어 인식에 비해 성능은 약간 떨어지지만, 확장성과 유연성 측면에서 유리한 가변어휘 인식이 대세를 이루고 있다. 그리고 실제 사용에서 사용자는 간투사나 주변잡음 (기침, 헤드셋 클릭, 침 삼킴 등)을 포함하여 발성하거나 비 숙련자의 경우 핵심어를 포함한 문장

책임저자: 김 민 제 (ink1978@mail.ulsan.ac.kr)
680-749 울산광역시 남구 무거2동 울산대학교
컴퓨터.정보통신공학
(전화: 052-259-1435)

단위로 발성하게 되고, 이것은 단어 인식기의 성능을 저하시키는 큰 요인이 된다. 따라서 이러한 문제를 해결하기 위하여 사용자가 발성한 음성에서 핵심어만을 인식하는 핵심어 검출 시스템이 필요하다. 핵심어 검출 시스템은 핵심 주제어만으로 의미가 통할 수 있는 응용분야에 효과적으로 활용될 수 있다 [1][2][3].

일반적으로 핵심어 검출은 인식하고자 하는 핵심어, 핵심어가 아닌 음성구간 (비핵심어), 그리고 묵음구간을 각각 모델링하고 이들의 연결 형태로 인식 network을 구성한다. 비핵심어 모델은 핵심어를 잡식하지 않으면서 비핵심어 음성부분을 효과적으로 표현할 수 있느냐에 따라 핵심어 검출 시스템의 성능이 크게 좌우되며 이에 대한 다양한 연구가 진행되어 왔다 [1][2][3][5][6][7].

비핵심어 모델을 구성하는 방법은 여러 가지가 있지만, 계산량과 인식성능을 종합적으로 고려하여 성능을 비교 실험한 결과, monophone 모델과 집단화 방법 (비핵심어 음성 부분 전체를 하나의 HMM으로 모델링)이 가장 효과적인 것으로 알려져 있다 [5][6].

Monophone을 이용한 비핵심어 모델링은 monophone을 몇 개의 group으로 clustering하여 비핵심어 모델로 구성하는 방법이다. 유사한 monophone을 grouping 하는 방법으로는 음향학적 음소분류 방법보다 우수한 성능을 나타내는 통계적인 방법을 주로 사용하고 있으며 통계적 유사도 측정을 위하여 weighted euclidean distance를 사용한 modified K-means알고리듬을 이용하여 clustering하는 방법을 적용하고 있다 [1]. 하지만 실험 결과 기존의 방법은 몇 가지 문제점을 보였다. Monophone 모델의 경우 monophone을 clustering하기 위해 사용하는 K-means 알고리듬은 초기 K개의 중심값과 거리측정 함수에 민감한 특징을 가지고 있으며, 각 state에 대한 특징을 반영하지 못하는 단점이 있다.

집단화에 의한 비핵심어 모델링은 문맥정보를 무시하고 임의적인 집단화 방법에 의해 다수의 가우시안 분포를 가지는 모델로 구성하는 방법이다. 이를 위해 비핵심어모델을 Gaussian Mixture Model (GMM)로 구현하고 있다. 이것은 비핵심어 음성구간 전체를 하나의 상태로 두어 multiple mixtures로 표현하는 방식이다 [2]. 집단화 방법은 훈련과정이 간단하지만 단어 전체를 하나의 모델로 구성하기 때문에 배경잡음이나 간투사처럼 음절의 특성을 가지는 부분을 잘 표현하지 못하며 single state로 구성함으로써 시간적 변위 특성을 잘 반영하지 못하는 단점이 있다.

따라서 본 논문에서는 이러한 문제점을 해결하기 위하여 두 가지의 비핵심어 모델링 방법을 제안하였다. 첫째는 monophone을 K-means알고리듬으로 clustering하는 방법을 개선하여 각 음소의 음향학적 정보와 유사도를 이용한 결정트리기반 상태 군집화 방법으로 상태를 tying하는 방법을 제안하였다. 두 번째는 음성구간 전체를 single state로 구성하는 방법을 개선하여 음절을 multi-state로 모델을 수정하고 mixture를 증가하면서 비핵심어 모델을 구성하였다.

본 논문의 구성은 다음과 같다. 2장에서는 Baseline 시스템의 구성에 대해 살펴보고, 3장에서는 본 논문에서 제안한 비핵심어 모델링 방법을 기술하였다. 그리고 4장에서는 실험 및 결과를 보이고, 5장에서는 결론 및 향후 연구계획에 대해서 언급하였다.

II. Baseline 시스템의 구성

본 연구에서 구성한 가변어휘 핵심어 검출 시스템은 그림 1과 같은 구조를 가지며, 비핵심어 모델링 방법의 성능을 비교하기 위하여 검출된 핵심어 후보들에서 오류를 제거하는 후처리 부분은 포함하지 않았다.

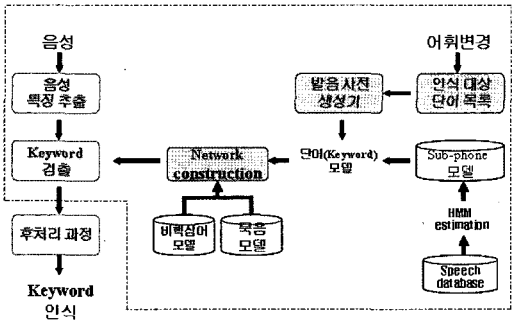


그림 1. 가변어휘 keyword spotting 시스템 구조
figure 1. Structure of Vocabulary-Independent Keyword Spotting System.

2.1. 음성 특징 추출

음성 특징 추출은 16kHz로 샘플링 된 신호를 12차 MFCC와 로그 에너지를 구하여 음성특징 파라미터로 사용하였다. 특징파라미터와 분석단위는 표 1과 같다.

표 1. 음성 특징 파라미터
Table 1. Feature parameter.

음성특징 파라미터	- MFCC 12차 + 로그 에너지 → 1차,2차 미분(총 39차) - 필터 뱅크 수 : 26 - Cepstral liftering 계수 : 22 - 분석 단위 : 20ms (10ms 중첩)

2.2. Subword 모델 구성

음향모델을 구성하는 단위는 단어, 음절, 음소, 트라이폰 등이 있으며 가변어휘 인식에서는 subword모델들의 결합으로 인식대상 어휘를 구성할 수 있는 음소나 트라이폰으로 구성되어진다.

음소 모델을 사용할 경우 적은 수의 모델로 구현함으로써 인식 소요 시간이 짧고 구현하기에 쉬운 반면에, 음향적 특성을 반영하지 못하므로 문맥 종속형 (Context - Dependent) 음소 (트라이폰)에 비해 인식성능이 떨어지는 단점을 갖고 있다.

트라이폰 모델을 사용할 경우 음향적 특성을 잘 표현할 수 있지만 신뢰도 높은 모델 파라미터를 추정하기 위해서는 방대한 양의 훈련용 데이터베이스가 필요하게 된다. 그러나 모든 한국어 음운 현상을 포함하는 DB를 구축하는 일은 현실적으로 어렵다. 따라서 모델의 신뢰도를 향상시키기 위한 기술이 필수적으로 요구된다 [8][9].

본 논문에서 음향모델은 3-state (시작, 종료 state제외)의 left- to-right 구조를 가지는 트라이폰 HMM을 사용하였으며, 음소는 한국어 음운현상을 반영하는 초성 (18개), 종성 (19), 종성 (7개), 묵음으로 구성된 45개의 음소 집합을 선정하였다. 그리고 모델에 대한 훈련데이터의 부족으로 인한 문제를 해결하기 위하여 유사한 통계적 특성을 갖는 모델의 state들을 하나의 그룹으로 묶어 상대적으로 모델의 신뢰도를 높이는 결정트리기반 상태 군집화 알고리듬을 사용하였다.

2.3. 인식 Network

인식 network은 그림 2와 같은 구조를 가지며 입력음성이 들어오면 인식과정을 통하여 하나의 핵심어를 검출하게 된다. 인식 network은 핵심어 모델, 비 핵심어 모델, 묵음모델의 결합으로 구성되며, 입력음성에는 핵심어가 반드시 하나만 존재한다고 가정하여 구성하였다. 그림 2에서 단어페널티는 입력음성에서 하나의 핵심어만을 검출하도록 하는 기능을 한다.

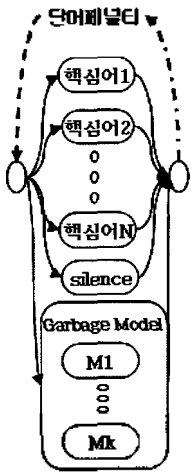


그림 2. 핵심어 검출 시스템의 인식 network
Figure 2. recognition network of keyword spotting system.

III. 비핵심어 모델 구성

핵심어 검출을 위해서 비 인식대상어휘, 주변 잡음, 간투사와 같은 핵심어가 아닌 음성을 모델링 할 필요가 있으며, 이를 비핵심어 모델이라고 한다. 따라서 핵심어 검출 시스템의 성능향상을 위하여 적절한 비핵심어 모델의 선택이 필요하다. 본 연구에서 핵심어 검출 시스템의 성능향상을 위해 제안하는 새로운 두 가지 방법의 비핵심어 모델링 방법은 다음과 같다.

3.1. 음소모델의 결정 트리기반 상태 군집화를 이용한 비핵심어 모델링

Monophone을 clustering하기 위하여 K-means 알고리듬을 이용하는 방법은 단순히 모든 state에서 weighted Euclidean distance합이 최소가 되도록 하는 방법이다. 이는 계산이 간단하며 각 음소상호간의 음향학적 지식이 필요하지 않다. 하지만 이 방법은 초기 K개의 중심과 거리측정 함수에 민감한 특징을 가지고 있으며, 단순히 각 모델의 모든 state에서 발생하는 평균과 분산을 이용하여 weighted euclidean distance를 구하게 되어 각 state에 대한 특징을 반영하지 못하게 된다.

따라서 이러한 문제점들을 해결하기 위하여 각 음소간의 음향학적 정보와 통계적인 정보를 모두 사용한 결정 트리기반 상태군집화를 이용한 비핵심어모델링 방법을 제안하였다. 이 방법은 각 음소의 음향학적 특징과 평가함수를 고려하여 각 모델의 state를 유사도에 따라 tying 시켜 각 state의 특징을 반영하였고, state간의 천이확률은 음소모델의 것을 그대로 사용하였다.

비핵심어 모델 구성 방법은 다음과 같은 단계를 거친다.

단계 1 : 음소 모델 구성 및 훈련

충분한 훈련데이터를 이용하여 각 음소에 대하여 3 state (시작, 종료 상태 제외)HMM모델을 구성한다. 이것은 기존의 문맥종속 모델로 구성하는 과정에서 얻을 수 있기 때문에 특별한 훈련 절차 없이 구성할 수 있다.

단계 2 : 음소에 대한 음향학적 분류 정의

각 음소에 대한 음향학적 특징을 고려하여 context question 구성한다. (ex. 유성음소, 무성음소 etc.)

단계 3 : 모든 음소모델을 각 상태별로 모음

모든 음소의 가우시안 분포를 상태별로 모은다.

단계 4 : 각 상태별 트리 구성

평가함수가 최대가 되게 하는 context question을 선택하여 두 개의 부분집합으로 나누어 가는 일련의 과정을 통해 각 상태별로 트리를 구성한다. 부분집합으로 분리했을 때 평가함수를 통한 관측 확률 값의 증가가 미리 정의한 임계치보다 작아지는 시점에서 분할을 멈추게 된다.

단계 5 : 상태 결합

트리가 구성되면 각 leaf node에 해당하는 음소들의 상태를 결합시킨다.

단계 6 : 재훈련

모든 상태가 결합되면 비핵심어 모델을 훈련데이터를 이용하여 가우시안 mixture를 순차적으로 증가시키면서 모델의 파라미터 재 추정을 통해 비핵심어모델을 구성하게 된다.

context question은 음향학적 음소 분류 특징을 고려하여 45개의 context question set을 구성하였으며, 평가함수는 수식 1과 같이 정의하였다 [10].

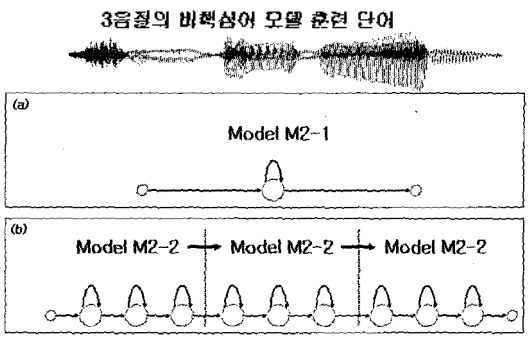
$$L=\sum_{s\in S}-\frac{1}{2}(n+\ln|(2\pi)^n|\Sigma_s|)\sum_{t=1}^{T}\gamma_s(t) \qquad (1)$$

여기서 Σs는 상태 s의 공분산이며, $\Sigma\gamma s\,(t)$는 상태 s가 발생하는 프레임 수이고, n은 특징벡터의 차수를 의미한다. 그리고 S는 현재 노드에 존재하는 상태 집합 ($s\in S$)을 나타낸다. 수식 1에서 공분산과 발생 수는 계산의 편리함을 위해 각 음소의 훈련 과정에서 구해진다.

3.2. 음절 단위의 multi-state 비핵심어 모델

기존의 방법은 음성구간 전체를 하나의 single state 모델로 구성한 다음 mixture의 수를 증가시키며 재훈련을 수행함으로써 비핵심어 모델을 구성하게 된다. 이 방법은 훈련과정은 간단하지만 단어 전체를 하나의 모델로 구성하였기 때문에 배경잡음이나 간투사처럼 음절의 특성을 가지는 부분을 잘 표현하지 못하며 시간적 변위 특성을 잘 반영하지 못하였다.

이러한 문제를 개선하기 위하여 그림 3과 같이 각 단어에서 음절을 하나의 모델로 구성하였다. 예를 들어 3음절 단어의 경우 기존 방법은 하나의 비핵심어 모델로 구성하지만, 음절 단위로 구성할 경우 동일한 3개의 비



(a) 기존 집단화 방법 (b) 음절기반 집단화 방법

그림 3. 훈련에서 비핵심어 모델 구성

figure 3. Construction of non-keyword model in training.

핵심어 모델의 연결로 구성되게 된다. 그리고 모델의 최소시간을 보장하기 위하여 state수를 1, 3, 9개로 multi-state single mixture를 가지는 모델을 구성하고 훈련시킨다. 그 후 모델의 mixture 개수를 하나씩 증가시켜 가면서 재훈련 과정을 반복하였다.

IV. 실험 및 결과

훈련과 성능평가는 ETRI 표준 한국어 공통음성 단어 DB를 훈련용 DB (Tr-DB)와 테스트용 DB (Te-DB)로 나누어 사용하였다 [4].

핵심어 모델은 남/녀 각각 500명이 발성한 Tr-DB를 사용하여 트라이폰 HMM을 구성하였고, Te-DB (남/녀 각각 100명이 각 10단어씩 발성한 2000단어)를 이용하여 가변어휘 단어인식 실험을 수행하였으며, 가장 좋은 99.13%의 인식률을 보이는 7개의 가우시안 mixture를 사용하였다.

비핵심어 모델 구성은 monophone 모델의 경우 핵심어 모델에서 트라이폰으로 변환하기 이전의 monophone 모델을 사용하였다. 그리고 집단화 모델의 경우 Tr-DB 중 핵심어 모델 구성에 사용되지 않은 남/녀 각각 100화자가 발성한 3,000단어를 이용하여 구성하였다.

비핵심어 모델에 따른 핵심어 검출기의 성능 평가를

표 2. 성능평가에 사용된 비핵심어 모델

figure 2. The model to be used at the performance test.

	기존모델	제안된 모델
monophone 모델	K-means기반 clustering 모델 (M-1) : cluster 수 1~10	결정 트리기반 상태 군집화 모델 (M-2) : mixture 수 1~10
집단화 방법을 이용한 GMM	음성구간 전체를 하나의 single state 모델 (G-1) : mixture 수 10,20,30	음절의 수에 따른 multi-state모델 (G-2) : mixture 수 10,20,30

위한 테스트는 표 2와 같은 4가지의 비핵심어 모델에 대하여 closed 테스트와 open 테스트를 수행하였다.

여기서 M-2모델은 state 수를 표 3과 같이 132개 (44음소)에서 13 (9.8%)개로 줄였고, 구성 결과 각 state에 따라 다른 음소들로 tying 되는 것을 확인하였다.

표 3. M-2모델의 leaf node
Table 3. leaf node of M-2 model.

state	group	음 소
2	s2_1	초성(ㄱ,ㄷ,ㅂ,ㅅ,ㅈ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ,ㄲ,ㄸ,ㅃ,ㅆ,ㅉ)
	s2_2	초성(ㄴ,ㄹ,ㅁ) 중성(ㅟ,ㅑ,ㅕ,ㅛ,ㅠ,ㅖ,ㅘ,ㅝ,ㅞ,ㅙ,ㅔ)
	s2_3	중성(ㅏ,ㅓ,ㅗ,ㅜ,ㅡ,ㅣ,ㅐ,ㅔ)
	s2_4	종성(ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅇ)
3	s3_1	초성(ㄲ,ㄴ,ㄹ,ㅁ) 중성(ㅟ,ㅑ,ㅕ,ㅛ,ㅠ,ㅖ,ㅘ,ㅝ,ㅞ,ㅙ,ㅔ,ㅏ,ㅓ,ㅗ,ㅜ,ㅡ,ㅣ,ㅐ,ㅔ)
	s3_2	종성(ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅇ)
	s3_3	초성(ㄱ,ㄷ,ㅂ,ㅅ,ㅈ,ㅊ,ㅋ,ㅌ,ㅍ,ㅎ,ㄸ,ㅃ,ㅆ,ㅉ,ㅂ)
4	s4_1	중성(ㅑ,ㅖ,ㅛ,ㅠ,ㅕ,ㅡ,ㅘ,ㅙ,ㅞ,ㅟ,ㅝ,ㅓ)
	s4_2	중성(ㅗ,ㅜ)
	s4_3	중성(ㅏ,ㅓ)
	s4_4	중성(ㅐ,ㅔ,ㅣ)
	s4_5	초성(ㅂ,ㅃ,ㅊ,ㄷ,ㄸ,ㄱ,ㄲ,ㅎ,ㅋ,ㅁ,ㄴ,ㅍ,ㄹ,ㅅ,ㅆ,ㅌ,ㅈ,ㅉ)
	s4_6	종성(ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅇ)

Closed 테스트는 Te-DB에서 표 4의 예와 같은 형태를 가지는 400문장 (남/녀 각각 100명, 어휘 수 100단어)에 대하여 수행하였다. 그리고 핵심어의 위치는 균등하게 분포되어 있다.

표 4. Closed 테스트 문장 구성의 예
Table 4. Example of closed test sentence.

■ "학교 *회사* 기차 자동차" (핵심어 회사)
■ "김병현 *박찬호* 이승엽 선동열" (핵심어 박찬호)
➡ 핵심어 1개와 비핵심어 3개가 연결된 구조
➡ 간투사 및 기타 잡음 미포함

표 5와 표 6은 closed 테스트 결과를 보여준다. 여기서 M-1모델은 cluster 수를 달리하였고, 나머지 모델들은 mixture 수를 순차적으로 증가하면서 구성한 모델에 대하여 테스트를 수행하였다.

표 5. Monophone 방법을 이용한 closed 테스트 결과
Table 5. Closed test result using monophone method.

개수	1	2	3	4	5	6	7	8	9	10
M-1 (cluster)	60.4	60.9	61.6	64.1	63.2	61.6	62.3	61.1	67.3	64.9
M-2 (mixture)	63.5	85.5	89.0	92.5	92.5	94.0	95.0	95.0	95.5	95.0

M-1과 M-2를 비교한 closed 테스트 결과 M-1은 cluster 수가 9개 일 때 가장 좋은 67.3%의 성능을 보였고, M-2는 mixture 수를 9개로 구성하였을 때 가장 좋은 95.5%의 성능을 보여 제안한 M-2 방법이 우수함을

표 6. GMM 방법을 이용한 closed 테스트 결과
Table 6. Closed test result using GMM method.

수 / mixture	G-1	G-2		
		1 state	3 state	9state
10	62.50	62.00	81.00	89.50
20	74.50	75.00	87.00	93.00
30	76.50	76.50	89.00	96.00

알 수 있다.

G-1과 G-2를 비교한 closed 테스트 결과, G-1은 20개의 mixture로 구성하였을 때 74.50%의 가장 높은 인식률을 보였고, G-2는 9 state로 모델링 하였을 경우 가장 좋은 96.00%의 인식률을 나타내었다. 따라서 closed 테스트에서 G-1보다 G-2를 9 state로 구성하는 것이 우수함을 알 수 있다.

Open 테스트는 4명의 화자가 각 100문장을 사무실 환경에서 발성한 400문장으로 실험하였다. 각 문장은 하나의 핵심어 (이름)을 포함하며 표 7은 테스트 문장 구성의 예를 제시하였다.

표 7. Open 테스트 문장 구성의 예
Table 7. Example of open test sentence.

■ "저 혹시 *박찬호*씨 있습니까?"
■ "거기 *이순신* 장군님 바꿔주세요"
➡ 기침, 입술, 한숨, 침넘기는 소리도 포함

표 8과 표 9는 closed 테스트와 동일한 방법으로 open 테스트를 수행한 결과를 보여준다.

표 8. Monophone 방법을 이용한 open 테스트 결과
Table 8. Open test result using monophone method.

개수	1	2	3	4	5	6	7	8	9	10
M-1 (cluster)	60.2	59.7	60.8	61.3	62.8	61.5	62.4	61.9	66.3	65.0
M-2 (mixture)	91.5	94.5	94.5	94.8	95.0	95.3	94.5	94.8	94.8	94.8

표 9. GMM 방법을 이용한 open 테스트 결과
Table 9. Open test result using GMM method.

수 / mixture	G-1	G-2		
		1 state	3 state	9state
10	94.25	94.50	96.25	92.96
20	94.25	94.25	95.25	92.98
30	94.25	95.25	95.75	92.95

M-1과 M-2를 비교한 open 테스트 결과 M-1은 cluster 수가 9개 일 때 가장 좋은 66.3%의 성능을 보였고, 제안한 M-2는 mixture 수를 6개로 구성하였을 때 가장 좋은 95.3%의 성능을 보여, M-2 방법이 29%의 인식률이 향상되어 우수함을 보였다.

G-1과 G-2를 비교한 closed 테스트 결과 G-1은 mixture 개수가 10개로 구성하였을 경우 가장 높은

94.25%의 인식률을 보였고, 본 논문에서 제안한 G-2는 3 state로 모델링 하였을 경우 가장 좋은 96.25%의 인식률을 나타내었다. 따라서 G-2가 우수한 비핵심어 모델링 방법이라는 결론을 얻었다.

제안한 M-2와 G-2 두 방법의 우수성을 비교하면 훈련과정의 효율성 측면에서는 핵심어 모델의 재활용이 가능한 monophone을 결정트리기반 상태 군집화를 이용하는 M-2가 우수하였지만, 핵심어 검출 성능 측면에서는 음절의 수에 따른 multi-state모델 (G-2)이 우수하였다.

V. 결론 및 향후 연구방향

본 논문에서는 트라이폰 HMM을 기반으로 하는 가변어휘 화자독립 핵심어 검출기의 성능을 향상시키기 위하여 새로운 비핵심어 모델링 방법을 제안하였다. 비핵심어 모델링 방법은 기존의 방법들의 문제점을 보완하여 각 음소의 음향학적 정보와 유사도를 이용한 decision tree 기반 state clustering방법으로 state를 tying하는 모델링 방법과 음절을 multi-state로 구성하는 방법을 제안하였다. 실험 결과 K-means 알고리듬을 이용한 monophone clustering 방법보다 결정트리기반 상태 군집화 방법을 사용하였을 경우 closed 테스트에서는 28%, open 테스트에서는 29%의 성능이 향상되었으며, 음성구간 전체를 single state 모델로 구성하는 방법보다 음절단위 multi-state 모델을 이용할 경우 closed 테스트에서는 2%, open 테스트에서는 22%의 인식률이 향상되어 기존의 방법보다 본 논문에서 제시한 비핵심어 모델링 방법이 우수한 결과를 보였다. 또한 제안한 두 알고리즘은 훈련과정의 편리함에서는 결정트리기반 상태군집화 방법이 우수하지만 인식결과에서는 음절단위 multi-state 모델을 이용한 방법이 우수하였다.

향후 연구에서는 핵심어 검출기의 성능을 향상시키기 위하여 본 논문에서 구성한 비핵심어 모델을 바탕으로 본 연구에서 구성되지 않은 후처리 과정을 추가하여 검출된 핵심어 후보들로부터 잘못 검출된 후보들을 효율적으로 제거하는 방안을 검토할 계획이다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT 연구센터 육성지원사업의 연구결과로 수행되었습니다.

참 고 문 헌

1. 황병한, "한국어 가변어휘 인식을 위한 음소 모델링 방법에 관한 연구", 부산대학교 석사졸업논문, 1999
2. 신영욱, "가변어휘 핵심어 검출 시스템의 구현 및 성능개선", 부산대학교 석사졸업논문, 2001
3. 김치수, 배건성, "고립단어 인식시스템에서 음성-비음성 식별에 관한 연구", 한국음향학회 학술대회지, 242-245, 1998.
4. 김상훈, 오승신, 정호영, 전형배, 김정세, "공통음성 DB 구축", 한국음향학회 학술대회지, 21-24, 2002
5. R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," ICASSP, 129-132, 1990
6. J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," IEEE Trans. Acoust., Speech, Signal Processing, 38 (11) 1870-1878, 1990
7. C.-H.Wu, Y.-J.Chen and G.-L.Yan , "Integration of phonetic and prosodic information for robust utterance verification", Vision, Image and Signal Processing, 147 55-61, 2000
8. Se-Jin Oh, Hyun-Yeol Chung, Cheol-Jun Hwang, Bum-Koog Kim, Ito, A., "New state clustering of hidden Markov network with Korean phonological rules for speech recognition", Multimedia Signal Processing, 39-44, 2001
9. Mei-Yuh Hwang, Xuedong Huang, Alleva, F.A., "Predicting unseen triphones with senones", Speech and Audio Processing, 4 (6) 412-419, 1996
10. Young S, Kershaw D, Odell J, Ollason D, Valtchev V, Woodland P, The HTK Book, Entropic Research Laboratories Inc., 1996

저자 약력

• **김 민 제 (Min-Je Kim)**

2004년: 울산대학교 컴퓨터.정보통신공학부 (학사)
2004년~2006년: 울산대학교 컴퓨터공학과 대학원 (석사)
※ 주관심분야: 음성인식, 자동음소분할

• **이 정 철 (Jung-Chul Lee)**
한국음향학회지 제 25권 5호 참조