2000년도 한국음향학회 학술발표대회 논문집 제 19 권 제 2(s) 호

가변어휘 핵심어 검출 시스템의 구현

신영욱, 송명규, 김형순
부산대학교 전자공학과

Implementation of Vocabulary-Independent Keyword Spotting System

Young Wook Shin, Myung Gyu Song, Hyung Soon Kim
Dept. of Electronics Engineering, Pusan National University
E-mail : {space73, mgsong, kimhs}@hyowon.cc.pusan.ac.kr

요 약

본 논문에서는 triphone 을 기본단위로 하는 HMM 에 의해 핵심어 모델을 구성하고, 사용자가 임의로 핵심어를 추가 및 변경할 수 있도록 가변어휘 핵심어 검출기를 구현하였다. 비핵심어 모델링 방법으로 monophone clustering 을 사용한 방법 및 GMM 을 사용한 방법의 성능을 비교하였다. 또한 후처리 과정에서 가변어휘 인식 구조에 적합한 anti-subword 모델을 사용하였으며 몇 가지 구현방식에 따른 후처리 성능을 검토하였다. 실험결과 비핵심어 모델로 monophone 을 clustering 하여 사용한 방법보다 GMM 을 사용한 경우 약간의 인식성능 개선을 얻을 수 있었으며, 후처리 과정에서 Kullback distance 를 이용한 anti-subword 모델링 방식이 다른 방식에 비해 우수한 결과를 나타냈다.

1. 서 론

음성인식은 입력 음성의 형태에 따라 크게 고립단어 인식과 연속음성인식으로 나눌 수 있다. 핵심어 검출은 자연스러운 연속음성으로부터 꼭 필요한 정보(keyword)를 추출해 내는 것으로 고립단어 인식이 지니는 발음상의 불편함과 연속음성인식 지니는 성능저조의 문제점을 모두 해결 할 수 있는 방식이다. 따라서 핵심 주제어만 검출해 내면 의미가 통할 수 있는 응용분야에 효과적으로 활용될 수 있다.

일반적으로 HMM 을 이용한 핵심어 검출은 인식하고자 하는 핵심어들, 핵심어가 아닌 음성부분 그리고 묵음구간을 각각의 HMM 으로 모델링하고 아무런 문법적 제한 없이 문장형태로 입력된 음성을 이들 HMM 들이 연결된 것으로 표현한다[1][2]. 여기서 비핵심어 모델이 핵심어 음성부분을 잠식하지 않으면서 비핵심어 음성부분 및 배경잡음 부분을 얼마만큼 효과적으로 표현해 줄 수 있느냐에 따라 핵심어 검출 시스템의 성능이 크게 좌우된다.

인식대상 어휘가 고정된 소규모 어휘인식의 경우 해당 어휘에 대한 음성 DB 를 이용하여 비교적 높은 인식성능을 얻을 수 있다. 그러나 만약 사용자가 인식하고자 하는 어휘를 변경하고자 할 경우 추가되는 어휘에 대한 음성 DB 의 재수집과 해당 모델의 재훈련을 해야 하는 불편함이 따르게 된다. 이에 반하여 가변어휘 인식은 인식대상 어휘를 임의로 추가 및 변경하더라도 미리 만들어 놓은 음소모델을 이용하여 인식대상 어휘 모델을 구성할 수 있으므로 변경된 부분에 대한 음성 DB 의 재수집이나 모델의 재훈련이 필요 없는 장점을 가진다.

본 논문에서는 triphone HMM 을 기반으로 하여 가변어휘 핵심어 검출기를 구현 하였다. 핵심어 검출기의 인식성능에 크게 영향을 미치는 비핵심어 모델의 구현방법으로 monophone 을 clustering 하여 사용하는 방식과 GMM 을 사용한 방식의 성능을 비교하였다. 이미 구해진 핵심어 후보들로부터 잘못 검출된 후보(false alarm)들을 효율적으로 제거하기 위한 후처리 과정에서는 가변어휘 환경에서 자동적으로 anti-keyword 모델의 생성이 가능하도록 anti-subword 모델을 이용하는 방식을 검토하였고 몇 가지 구현방식에 따른 성능을 비교하였다.

2. 가변어휘 핵심어 검출 시스템의 구성

본 논문에서 구현한 가변어휘 핵심어 검출기의 구조는 그림 1 과 같다. 다양한 음운현상이 반영된 음성 DB 로부터 triphone HMM 을 훈련한다. 인식해야 할 핵심어가 정해지면 발음 표기 변환을 통하여 인식대상 어휘를 음소열로 변환 시킨 뒤 해당되는 음소 모델들을 연결하여 핵심어 모델을 구성한다. 핵심어 검출 단계에서 미지의 음성이 들어 오면 핵심어 모델 및 비핵심어 모델, 묵음 모델의 네트웍으로 구성된 연결단어 인식과정을 통해 핵심어를 찾아내게 된다. 본 논문에서는 입력음성에 핵심어가 하나만 들어 있다는 가정하에 인식 네트웍을 구성하였다. 후처리 과정에서는 검출된

핵심어의 신뢰도를 조사하여 핵심어로 판단하기 곤란한 것들을 제외시킴으로써 오인식에 따른 문제를 줄이도록 하였다.

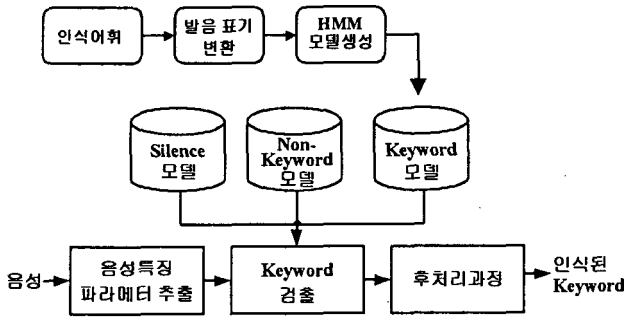

그림 1. 가변어휘 keyword spotting 시스템의 구성도

음성특징 파리미터 추출은 음성신호를 16kHz 로 샘플링하여 20msec 프레임 단위로 10msec 씩 shift 하면서 전달함수가 $1\text{-}0.97z^{-1}$인 디지털 필터로 preemphasis 를 하고, 여기에 다시 Hamming window 를 씌운 후 12 차 MFCC 를 구하고, 음성신호의 시간축 상에서의 변화특성 정보를 보존하기 위하여 delta 파라미터를 구하여 총 24 차 특징벡터를 사용하였다.

triphone 모델은 제한된 훈련 DB 에 대해 모델 파라미터 추정의 신뢰도를 높일 수 있도록 tree based clustering 을 사용하여 left-to-right 연속확률분포를 가지는 tied state HMM 으로 훈련하였다. 모델당 state 는 3 개를 사용하였으며 state 당 mixture 개수에 따른 인식실험을 하였다. 그리고, 묵음 모델은 10 개의 state 를 가지되 state 당 하나의 mixture 를 가지도록 훈련 하였다.

3. 비핵심어 모델의 구성

Filler model 을 이용하는 핵심어 검출 방식에 있어서 비핵심어 모델은 핵심어와 핵심어가 아닌 음성을 구분지어 주는 역할을 한다. 따라서 인식성능의 향상을 위해서는 적절한 비핵심어 모델의 선택이 필요한데, 본 논문에서는 비핵심어 모델링 방법으로 monophone 들을 clustering 하여 사용한 방식과 GMM 을 사용한 방식을 검토하였다.

3.1 Monophone cluster 비핵심어 모델

유사한 특성의 음소를 grouping 하는 방법으로 음성학적 지식을 이용하는 방법과 통계적인 방법이 사용될 수 있다[7]. 본 논문에서는 상대적으로 우수한 성능을 가지는 통계적 방법에 의한 monophone clustering 방법을 사용하였다. 45 개의 monophone 모델을 훈련한 후 각각의 모델들의 확률분포로부터 서로간의 거리를 구한 후, modified k-means(MKM) 알고리즘을 사용하여 grouping 을 하였다[4]. 단일 mixture 를 가지는 음소 모델들 사이의 거리 척도는 다음과 같이 정의하였다.

$$D(p_i,p_j)=\sum_{d=1}^{N}D_d(p_i,p_j) \qquad (1)$$

여기서 p_i, p_j는 각각 i 와 j 번째 음소를 나타내고, N은 음소모델의 state 수를 나타낸다. $D_d(p_i,p_j)$는 분포간의 거리로서 다음 식과 같이 주어진다.

$$D_d(p_i,p_j)=\frac{1}{V}\sum_{k=1}^{V}\frac{(\mu_{idk}-\mu_{jdk})^2}{\sigma_{idk}\sigma_{jdk}} \qquad (2)$$

여기서, V는 음성특징벡터의 차원이고 μ_{idk}, μ_{jdk}, $\sigma_{idk}, \sigma_{jdk}$는 각각 i 번째 및 j 번째 음소의 d 번째 분포에서 k 번째 파라메타의 평균 및 표준편차이다.

3.2 GMM 을 이용한 비핵심어 모델의 구성

비핵심어 모델을 구성하는 또 다른 방법으로 GMM 을 사용하는 방식을 검토해 보았다. 이것은 single state multiple mixture 를 사용하는 것인데, 먼저 훈련 DB 전체를 이용하여 single state single mixture 를 가지는 HMM 을 만든 후 state 의 mixture 개수를 하나씩 증가 시키면서 재훈련하는 과정을 반복하여 원하는 개수가 될 때까지 만든 것이다.

4. 후처리 방식 검토

본 논문에서는 후처리 방법으로 filler 모델을 이용한 likelihood ratio scoring 방법과 가변어휘 상황에서 자동으로 anti-keyword 를 구성하여 후처리가 가능하도록 anti-subword 모델을 구성하는 방식을 검토하였다.

4.1 Log likelihood ratio scoring 방법[2]

이 방법은 핵심어라고 찾아진 구간의 likelihood 와 이 구간을 다시 비핵심어와 묵음 모델만으로 구성된 network 에 통과시켜 얻은 likelihood 의 차이를 이용하는 것으로서 비핵심어와 묵음 모델의 확률에 비해 핵심어 모델에서의 확률이 얼마나 높은가 하는 점을 판단 기준으로 하는 방법이다. 이것을 식으로 나타내면 다음과 같다.

$$S_w=\log P(O_{t_s}^{t_e}\mid w)-\log P(O_{t_s}^{t_e}\mid f) \qquad (3)$$

여기서 w와 f는 각각 핵심어 및 비핵심어와 묵음 모델을 의미하고 $O_{t_s}^{t_e}$는 프레임 t_s로부터 t_e까지의 관찰벡터 파리미터 열을 나타낸다. 이 score 를 적절한 임계치와 비교하여 핵심어 여부를 최종적으로 판단한다.

4.2 Anti-subword model 을 이용한 방법[5]

subword 모델에 대한 anti-subword 모델을 만듦으로써

가변어휘 상황에서 anti-keyword 를 자동으로 구성하여 발화검증을 하는 방식이 사용되어지고 있다. 본 논문에서는 45 개의 monophone 에 대한 anti-subword 모델을 구성한 후 이것들을 연결함으로써 가변어휘에 핵심어에 대한 anti-keyword 를 구성하였다. anti-subword 모델을 만드는 방법으로 3 가지 방식을 검토하였다.

첫 번째는 분포간의 거리척도로서 Kullback distance[6]를 이용하여 특정 monophone 에 대해 가장 혼동 가능성이 높은 모델 하나를 그 monophone 에 대한 anti-subword 모델로 사용하는 방식이다. 두 분포 $f(x)$ 및 $g(x)$ 에 대한 Kullback distance 는 다음과 같이 주어진다.

$$KL(f,g) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx \qquad (4)$$

두 번째는 첫 번째 방식에서 혼동 가능성이 가장 높은 상위 2 개를 이용하여 anti-subword 를 구성하는 방식이다. anti-subword 모델의 천이확률은 reference monophone 의 것을 그대로 가져다 쓰고 state 의 분포는 구해둔 두 모델의 해당 state 의 분포를 mixture 로 사용하되 weight 값은 동등하게 주었다.

세 번째는 첫 번째 방법에서와 같이 가장 혼동가능성이 높은 것 하나를 구하되 Kullback distance 대신 수식 (1)을 이용한 방식이다.

5. 실험 및 결과

핵심어 검출 실험으로 인식대상 어휘가가 고정된 경우 및 가변어휘인 경우에 대한 실험을 하였다.

고정어휘 핵심어 검출의 경우 한국전자통신연구소에서 구축한 부서명 DB 를 사용하였다. 남자 50 명이 22 개의 혼동 가능성이 있는 부서명을 단어 및 문장형태로 발성한 DB 중에서 35 명분을 훈련에 사용하였고, 15 명분을 테스트에 사용하였다.

가변어휘 핵심어 검출 실험의 경우는 음소모델의 훈련을 위해 한국전자통신연구소에서 구축한 음소열 최적화 단어 DB(POW 3848 DB [3]) 중 일부를 사용하였다. 테스트는 부서명 DB 중에서 남자 50 명이 부서명 22 개를 발음한 단어 및 문장형태 음성 DB 를 사용하였다.

실험결과를 표 1 에 나타내었다. 표 1 에서 비핵심어 모델로 GMM 을 사용한 경우 mixture 개수를 15 개에서 22 개까지 사용하여 실험 한 것 중에서 성능이 가장 좋은 경우에 대한 결과를 나타내었고, monophone cluster 를 사용한 경우는 cluster 개수를 3 개에서 10 개까지 사용하여 실험을 한 것 중에 가장 좋은 경우만을 나타내었다. 각각의 경우에 대한 GMM 의 해당 mixture 수와 monophone cluster 개수를 비핵심어 모델 칸에 표시하였다.

실험 결과에서 문장인식 결과를 보면, 고정어휘의 경우 두 가지 비핵심어 모델링 방법에 따른 인식성능의 큰 차이는 나타나지 않았고, 제한된 양의 훈련 DB 를 이용해도 우수한 성능을 얻을 수 있었다.

가변어휘의 경우는 비핵심어 모델로 GMM 을 사용한 방식이 monophone 을 clustering 하여 사용한 방법보다 조금 더 우수한 성능을 나타냈고, 고정어휘 인식의 최고 성능에 비해서는 약 4% 정도의 인식률이 저하된 결과를 얻었다.

또한 문장형태 입력음성에 대해 핵심어가 입력음성의 시작부분에 위치하는 경우 및 중간부분에 위치하는 경우로 나누어 실험을 각각 해 보았다. 그 결과 핵심어가 문두에 오는 경우는 97.3%, 문중에 오는 경우는 94.0%의 인식성능을 각각 얻을 수 있었다.

GMM 을 사용하여 최고 성능을 얻은 결과에서 문장 입력 형태에 대한 오인식의 유형을 분석해본 결과 핵심어 구간을 잘 못 검출한 경우에 의한 것이 80% 정도였고, 핵심어 구간을 정확히 검출했으나 혼동가능성이 높은 일부 단어들에 의한 substitution error 가 그 나머지를 차지 했다.

표 1. 핵심어 검출 실험 결과

인식 형태		비핵심어모델		# of tied state mixture			
				1	3	5	6
고정어휘	문장	GMM	20	97.9	97.5	97.5	**98.3**
		Mono	8	97.5	97.9	97.9	98.3
	단어	GMM	15	99.7	100	100	100
		Mono	3	99.7	100	100	100
가변어휘	문장	GMM	20	89.7	92.2	93.6	**94.6**
		Mono	6	89.1	91.6	92.9	93.0
	단어	GMM	15	98.5	99.4	99.6	99.6
		Mono	5	98.4	99.4	99.5	99.5

후처리 실험은 가변어휘 핵심어 검출에서 비핵심어 모델로는 20 개 mixture 를 가지는 GMM 을 사용하고, tied state 의 mixture 를 6 개로 사용한 경우에 대해 실험을 하였다. 핵심어 기각률에 대한 핵심어 검출률을 그림 2 에 나타내었다. 실험 결과 Kullback distance 를 이용하여 가장 유사한 모델 하나를 anti-subword 로 사용한 경우가 가장 우수한 성능을 나타냈다. 이 경우에 대해서 핵심어의 duration 으로 추가적인 정규화를 했을 때와 하지 않았을 때의 성능을 그림 3 에 나타내었는데 성능차이는 거의 나타나지 않았다.

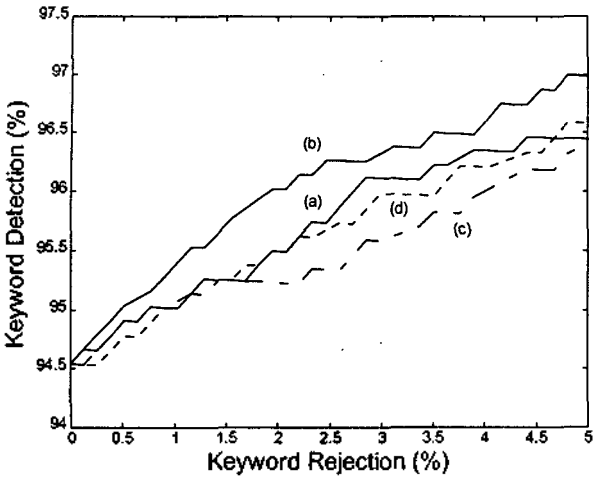

그림 2. 후처리 방법에 따른 성능비교
(a) Mixture 20 개를 가지는 GMM 을 비핵심어 모델로 사용한 경우.
(b) Kullback distance 를 이용하여 가장 유사한 subword 모델을 해당 anti-subword 모델로 사용한 경우.
(c) Kullback distance 를 이용하여 가장 유사한 상위 2 개를 이용하여 anti-subword 모델을 구성한 경우.
(d) (b)의 방법과 동일하되 수식 (1)의 거리척도를 이용한 경우

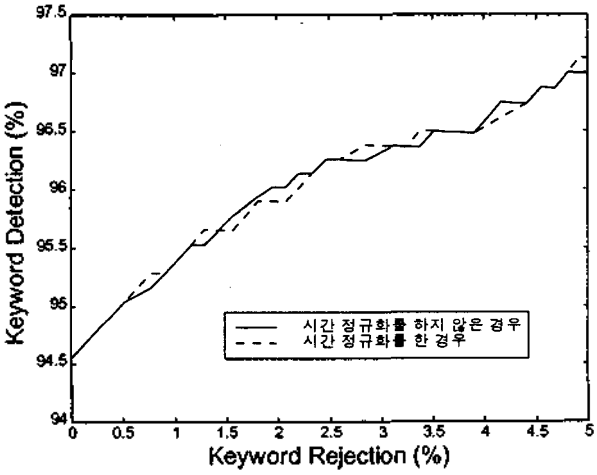

그림 3. Kullback distance 를 이용한 방법에서 그림 2 의 (b)방법에 대해 시간 정규화를 하지 않은 경우와 시간 정규화를 한 경우의 성능차이.

6. 결 론

본 논문에서는 triphone HMM 을 기반으로 인식하고자 하는 핵심어를 임의로 추가 및 삭제할 수 있는 가변어휘 핵심어 검출기를 구현하였다. 핵심어 검출기의 성능 개선과 관련해서 두 가지 비핵심어 모델링 방법을 검토한 결과 가변어휘 상황에서는 GMM 을 사용한 방법이 monophone 을 clustering 하여 사용한 방법보다 조금 더 나은 성능을 보였다. GMM 비핵심어 모델을 사용하여 문장입력에 대해 94.6%, 단어입력에 대해 99.6%의 인식 성능을 얻을 수 있었다.

후처리 과정에서 anti-subword 모델을 구성함으로써 가변어휘 상황에서도 anti-keyword 를 구성할 수 있도록 하였으며, Kullback distance 를 이용한 anti-subword 모델링 방식이 다른 방법에 비해 성능이 우수하게 나타났다. 후처리 과정을 통하여 핵심어 기각률이 3%일 때 96.4%의 인식성능을 얻을 수 있었다. 앞으로 핵심어 검출 시스템의 성능을 보다 향상시키기 위해서는 핵심어 및 비핵심어 모델들을 효과적으로 모델링 하는 방법이 연구되어야 할 것으로 판단되며 현재 핵심어 검출기의 후처리 과정에서 anti-subword 모델링에 대한 추가적인 연구가 진행되고 있다.

참 고 문 헌

[1] 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, vol.38, no.11, pp.1870-1878, Nov. 1990.

[2] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in Proc. IEEE ICASSP, pp.129-132,1990.

[3] Y. J. Lim and Y. J. Lee, "Implementation of the POW (phonetically optimized words) Algorithm for Speech Database," in Proc. IEEE ICASSP, vol.1, pp.89-92, May 1995.

[4] J. G. Wilpon and L. R. Rabiner, "A modified K-Means clustering algorithm for use in isolated word recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.33, no.3, pp.587-594, June 1985.

[5] R. A. Sukkar and C. H. Lee ,"Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," IEEE Trans. Speech. Audio Processing, vol.4, no 6, pp.420-429, Nov. 1996.

[6] S. Kullback and R. Leibler. "On the information and sufficiency," Annals of Mathematical Statistics, vol.22, pp.79-86, 1951.

[7] 이황림, 김형순 외, "음소 HMM 을 이용한 핵심어 검출 시스템의 성능향상에 관한 연구," 한국음향학회지 제 16 권 제 8 호, pp.60-67, 1997 년 8 월