가변어휘 핵심어 검출을 위한 비핵심어 모델링 및 후처리 성능평가*

Performance Evaluation of Nonkeyword Modeling and Postprocessing for Vocabulary-independent Keyword Spotting

김 형 순** · 김 영 국** · 신 영 욱**
Hyung Soon Kim · Young Kuk Kim · Young Wook Shin

ABSTRACT

In this paper, we develop a keyword spotting system using vocabulary-independent speech recognition technique, and investigate several non-keyword modeling and post-processing methods to improve its performance. In order to model non-keyword speech segments, monophone clustering and Gaussian Mixture Model (GMM) are considered. We employ likelihood ratio scoring method for the post-processing schemes to verify the recognition results, and filler models, anti-subword models and N-best decoding results are considered as an alternative hypothesis for likelihood ratio scoring. We also examine different methods to construct anti-subword models. We evaluate the performance of our system on the automatic telephone exchange service task. The results show that GMM-based non-keyword modeling yields better performance than that using monophone clustering. According to the post-processing experiment, the method using anti-keyword model based on Kullback-Leibler distance and N-best decoding method show better performance than other methods, and we could reduce more than 50% of keyword recognition errors with keyword rejection rate of 5%.

Keywords: Speech Recognition, Keyword Spotting, Vocabulary-independent, Non-keyword Model, Postprocessing

* 본 논문은 한국과학재단 목적기초연구(R01-2000-000-00275-0) 지원으로 수행되었음.

** 부산대학교 전자공학과

1. 서 론

음성인식은 입력 음성의 형태에 따라 크게 고립단어인식과 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[1]. 따라서 핵심 주제어만 검출해 내면 의미가 통할 수 있는 응용분야에 효과적으로 활용될 수 있다.

일반적으로 HMM을 이용한 핵심어 검출은 인식하고자 하는 핵심어들, 핵심어가 아닌 음성부분 그리고 묵음구간을 각각의 HMM으로 모델링하고 아무런 문법적 제한 없이 문장형태로 입력된 음성을 이들 HMM들이 연결된 것으로 표현한다[1]-[3]. 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. 그리고 핵심어 검출기의 인식결과로 나온 후보들에 대해 적절한 방법으로 신뢰도를 평가하여 잘못 검출된 후보들을 효율적으로 제거하는 후처리 과정을 둠으로써 오인식에 따른 문제를 완화시킬 수 있다.

본 논문에서는 인식하고자 하는 핵심어를 자유롭게 추가 및 변경할 수 있는 가변어휘 핵심어 검출기를 구현하였다. 이를 위해 다양한 음운현상이 반영된 음성 데이터베이스를 이용하여 음소모델을 미리 구성하고, 인식대상 핵심어 모델은 발음사전에 따라 음소모델을 연결하여 구성하는 가변어휘 음성인식 기술을 사용하였다. 그리고, 본 논문에서는 가변어휘 핵심어 검출 시스템의 성능에 중요한 영향을 미치는 비핵심어 모델링 방법과 후처리 방법에 대해 몇 가지 접근 방법을 적용하여 그 성능을 비교하였다. 비핵심어 모델의 구성 방법으로 monophone을 clustering하여 사용하는 방식과 Gaussian mixture model (GMM)을 사용하는 방식을 검토하였다. 또한 후처리 과정에서는 log likelihood ratio (LLR) scoring에 기반한 방식으로 filler 모델을 이용하는 방식, anti-keyword 모델을 이용하는 방식, 그리고 N-best decoding을 이용하는 방식을 검토하였다. 가변어휘 환경에서 자동적으로 anti-keyword 모델의 생성이 가능하도록 하기 위하여 anti-subword 모델을 구축하여 사용하였다.

본 논문의 구성은 다음과 같다. 서론에 이어 2 장에서 가변어휘 핵심어 검출 시스템에 관해 설명한다. 그리고 3 장과 4 장에서 본 논문에서 검토한 비핵심어 모델링 방식과 후처리 방식들에 대해 각각 설명한다. 이어서 5 장에서 실험환경 및 결과를 기술한 후, 마지막으로 6 장에서 결론을 맺는다.

2. 가변어휘 핵심어 검출 시스템의 구성

본 논문에서는 가변어휘 음성인식 기술을 이용하여 사용자가 임의로 핵심어를 추가 및

변경할 수 있는 가변어휘 핵심어 검출기를 그림 1과 같이 구성하였으며, 이를 설명하면 다음과 같다. 먼저 가변어휘 인식을 위해 다양한 음운현상이 반영된 음성 DB로부터 triphone HMM을 훈련한다. 인식해야 할 핵심어가 정해지면 발음 표기 변환을 통하여 인식대상 어휘를 음소열로 변환시킨 다음, 해당되는 음소 모델들을 연결하여 핵심어 모델을 구성한다 핵심어 검출 단계에서 미지의 음성이 들어오면 핵심어 모델 및 비핵심어 모델, 묵음 모델의 네트웍으로 구성된 연결단어 인식과정을 통해 핵심어를 찾아내게 된다. 본 논문에서는 입력음성에 핵심어가 하나만 들어 있다는 가정 하에 인식 네트웍을 그림 2와 같이 구성하였다. 후처리 과정에서는 검출된 핵심어의 신뢰도를 평가하여 핵심어로 판단하기 곤란한 것들을 기각시킴으로써 오인식에 따른 문제를 줄이도록 하였다.

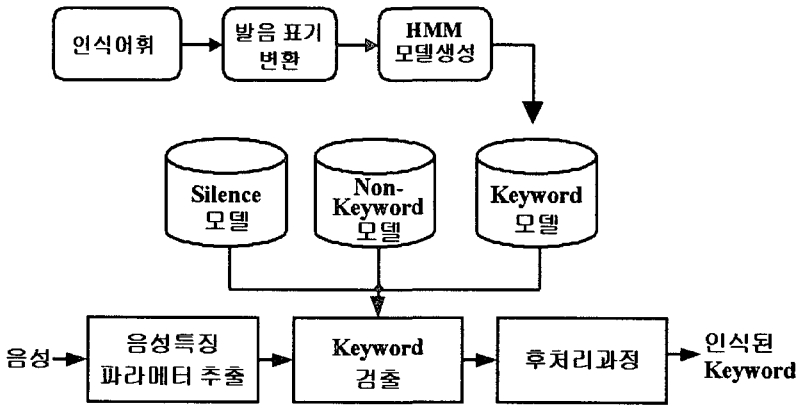

그림 1. 가변어휘 핵심어 검출 시스템의 구성도

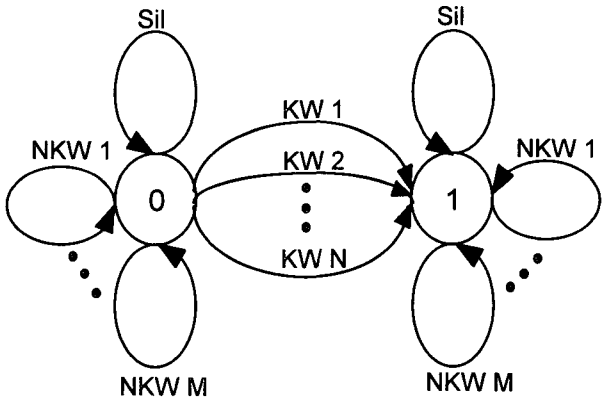

그림 2. 문장 당 한 개의 핵심어가 있을 때의 인식 네트웍의 예

3. 비핵심어 모델의 구성

핵심어 검출 시스템에서 핵심어에 해당되지 않는 음성(non-keyword) 구간에 대한 모델과 비음성 구간에 대한 모델을 묶어서 filler 모델이라고 부른다. 실제로 filler 모델을 어떻게 정의하고 구현할 것인가 하는 점이 핵심어 검출 알고리즘 사이의 가장 큰 차이점으로 부각되는데, 이 filler 모델이 핵심어 부분을 잠식하지 않으면서 비핵심어 음성 부분 및 배경잡음 부분을 얼마만큼 효과적으로 표현해 줄 수 있는가에 따라 핵심어 검출 시스템의 성능이 크게 좌우된다. 본 논문에서는 비핵심어 모델링 방법으로 monophone 모델들을 통계적으로 clustering하여 사용하는 방식과 비핵심어 음성 구간을 단일상태의 복수 mixture로 표현되는 Gaussian mixture model (GMM)을 사용한 방식을 각각 검토하였다

3.1. Monophone cluster 비핵심어 모델

모든 monophone을 그대로 사용할 경우 핵심어 부분을 잠식할 우려가 있고, 또 하나로 묶어서 사용할 경우 핵심어가 아닌 부분을 제대로 표현해 주지 못할 수 있기 때문에 몇 개의 group으로 clustering을 하여 비핵심어 모델로 사용하였다. 이 방식은 묵음을 제외한 45 개의 monophone에 대해 HMM 모델을 훈련한 후 훈련된 monophone 모델끼리의 거리를 구하여 그 거리가 가까운 것끼리 묶어서 몇 개의 group으로 grouping을 하는데, 유사한 음소들을 grouping하는 방법[4]으로 음성학적 지식을 이용하는 방법과 통계적인 방법이 사용될 수 있다. 본 논문에서는 상대적으로 우수한 성능을 가지는 통계적인 방법에 의해 monophone 모델을 clustering하였다. Monophone HMM은 3 개의 상태를 가지고 상태당 1 개의 mixture를 가지는 left-to-right 구조의 continuous density HMM으로 구성을 하였다.

3.1.1. Weighted Euclidean distance에 의한 clustering

이 방식은 monophone 모델의 확률분포로부터 모델들끼리의 거리를 구할 때 weighted Euclidean distance를 사용하여 거리를 구하고, modified k-means (MKM) [5] 알고리즘에 의해서 monophone 모델들을 몇 개의 그룹으로 나눈 방식이다.이때 음소모델들 사이의 거리척도는 다음과 같이 주어진다.

$$D_{WE}(p_i, p_j) = \sum_{s=1}^{N} D_s(p_i, p_j) \tag{1}$$

여기서 p_i , p_j는 각각 i와 j번째 음소를 나타내고, N은 음소모델의 상태수를 나타낸다. $D_s(p_i, p_j)$는 상태간의 거리로써 다음과 같이 주어진다.

$$D_s(p_i, p_j) = \frac{1}{V}\sum_{d=1}^{V}\frac{(\mu_{isd} - \mu_{jsd})^2}{\sigma_{isd}\sigma_{jsd}} \tag{2}$$

여기서 V는 음성 특징벡터의 차원이고 μ_{isd}, μ_{jsd}, σ_{isd}, σ_{jsd}는 각각 i 번째 및 j 번째 음소의 s 번째 상태의 d 차원의 평균 및 표준편차이다.

3.1.2. Kullback-Leibler distance에 의한 clustering

확률분포 간의 거리를 구할 때 사용하는 거리척도들의 성능을 비교한 한 연구 결과에 따르면, Kullback-Leibler(KL) distance에 기반을 둔 방식이 Euclidean distance와 Mahalanobis distance를 사용한 방식보다 성능면에서 우수하다고 보고 되었다[6]. 본 논문에서도 monophone 모델들끼리의 거리를 구할 때 모델의 확률분포간의 거리척도로써 Kullback-Leibler distance[7]를 사용한 방식을 검토하였다. 랜덤변수 X에 대한 두 개의 확률분포 $f(x)$, $g(x)$ 사이의 Kullback-Leibler distance는 다음과 같이 주어진다.

$$KL(f(x),g(x)) = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx \tag{3}$$

상태의 관찰 확률분포로 Gaussian 분포를 사용할 경우 두 분포간의 거리는 다음의 결과식으로 정리된다.

$$KL\left(f_X(x),f_Y(x)\right) = -\frac{1}{2}\left[\ln\left(\frac{\sigma_X^2}{\sigma_Y^2}\right) - \frac{\sigma_X^2+(m_X-m_Y)^2}{\sigma_Y^2} + 1\right] \tag{4}$$

여기서 m_X, σ_X^2 및 m_Y, σ_Y^2는 각각 확률분포 $f_X(x)$ 및 $f_Y(x)$의 평균과 분산이다. 그런데 Kullback-Leibler distacne는 대칭적(symmetric)이지 못한 성질을 가지고 있다. 즉, 비교대상 모델의 순서가 바뀌면 거리의 결과값이 다르게 나오게 되어 있다. 따라서 본 논문에서는 두 모델에 대해 대칭적이 되도록 하기 위하여 다음과 같이 변형된 수식을 사용하였다[7].

$$KL2(f(x),g(x)) = KL2(g(x),f(x)) = \frac{1}{2}\left[KL(f(x),g(x)) + KL(g(x),f(x))\right] \tag{5}$$

이러한 거리척도를 이용하여 모델간의 거리는 다음과 같이 구하였다.

$$D_{KL}(p_i,p_j) = \sum_{s=1}^{N}\sum_{d=1}^{V} KL2\left(f_{isd}(x),g_{isd}(x)\right) \tag{6}$$

여기서 N은 모델의 상태수를 나타내고 V는 음성 특징벡터의 차원을 나타내며 $f_{isd}(x)$는 음소 p_i의 s 번째 상태의 d 차원의 확률분포를 나타낸다. 수식 (6)를 이용하여 monophone

모델들 사이의 거리를 구하고 3.1.1 절과 동일하게 modified k-means 알고리즘을 사용하여 clustering을 하였다.

3.2. GMM을 이용한 비핵심어 모델

비핵심어 모델을 구성하는 또 다른 방법으로 Gaussian mixture model (GMM)을 사용하는 방식을 검토하였다. 이는 비핵심어 음성 구간 전체를 단일 상태의 복수 mixture를 가지는 Hidden Markov Model (HMM)로 표현하는 것과 동일하다. GMM은 그림 3과 같이 복수 개의 Gaussian density의 가중합(weighted sum)으로 구성된다[8].

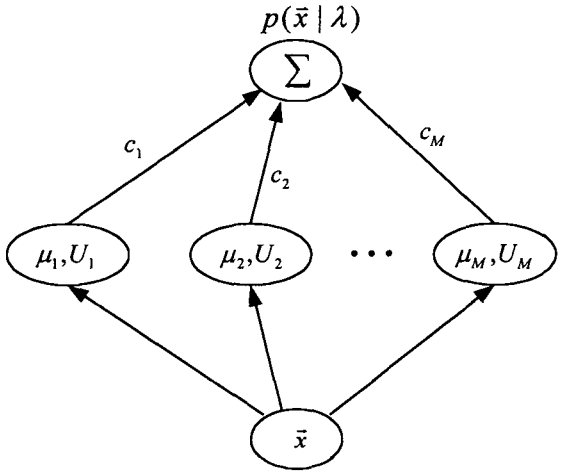

그림 3. M 개의 mixture를 가지는 GMM

이것을 수식으로 나타내면 다음과 같다.

$$p(\vec{x}\mid\lambda)=\sum_{i=1}^{M}c_i N_i(\vec{x},\mu_i,U_i) \tag{7}$$

여기서 $\vec{x}$는 음성 특징벡터이고, c_i는 i번째 mixture의 가중치(weight)이고, 모든 가중치를 더하면 1이 된다. N_i는 모델 λ의 i번째 mixture의 Gaussain 확률분포로 평균이 μ_i이고 공분산이 U_i이다. 본 논문에서 사용한 GMM의 훈련과정은 다음과 같다. 먼저 묵음구간을 제외한 훈련데이터 전체로부터 먼저 단일 상태 및 단일 mixture를 가지는 모델을 만든다. 그 이후 이 모델의 mixture 개수를 하나씩 증가시켜 가면서 Expectation-Maximization (EM) 알고리즘에 의한 재훈련 과정을 반복한다[8]. 이렇게 원하는 mixture 개수가 될 때까지 훈련하여 비핵심어 모델을 구성하였다.

4. 후처리 방식 검토

핵심어 검출 시스템에서 후처리 과정을 수행하는 목적은 핵심어 검출의 시스템에서의 오류를 감소시켜 그 성능을 보다 향상시키기 위한 것이다. 이것은 핵심어 검출의 성능이 완벽하지 못한 상태에서 실제 응용분야에 적용하기 위해서는 잘못된 결과를 출력시키기보다는 결과를 출력시키지 않고 재시도를 하거나 경우에 따라서는 인식을 포기하는 것이 많은 경우 문제의 소지를 줄일 수 있다는 판단에 근거를 둔 것이다. 실제로 본 논문의 task domain으로 선정된 전화교환업무를 예로 들어 설명하면, 입력음성으로부터 잘못된 핵심어를 검출하여 다른 부서로 전화연결을 하는 것보다는 핵심어를 검출해 내지 못했음을 통보하고 다시 한번 발음해 달라고 요구하는 것이 사용자의 불편을 상대적으로 감소시킬 수 있다. 본 논문에서는 후처리 방법으로 filler 모델을 이용한 likelihood ratio scoring 방법과 가변어휘 상황에서 자동으로 anti-keyword를 구성하여 후처리가 가능하도록 anti-subword 모델, 그리고 N-best decoding을 이용하는 방식을 검토하였고, 후처리 방법의 신뢰도의 정규화 방법으로는 시간 정규화 이외에 log likelihood의 절대값에 의한 정규화 방식을 적용하였다.

4.1. Filler 모델을 이용한 방법

이 방법은 핵심어라고 검출된 구간에서의 확률값이 filler 모델의 확률에 비해 얼마나 높은가 하는 점을 판단기준으로 하는 것이다[3]. 입력음성으로부터 핵심어 후보 및 이 핵심어가 존재하는 음성 구간 정보를 검출해 내고, 핵심어라고 찾아진 구간을 다시 비핵심어 모델과 묵음 모델 만으로 구성된 filler 네트웍에 통과시켜 filler 모델의 likelihood를 구하게 된다. 전체 네트웍에 의해 특정 핵심어가 프레임 t_s로부터 프레임 t_e까지의 구간에서 검출되었다고 하면, 인식된 핵심어의 log likelihood로부터 이 구간에 해당하는 filler 모델의 log likelihood를 뺀 값을 likelihood ratio score S_w라고 하고 다음 수식과 같이 나타낸다.

$$S_{w_k} = \frac{1}{t_e - t_s}\left[\log P\left(O_{t_s}^{t_e} | W_k\right) - \log P\left(O_{t_s}^{t_e} | F\right)\right] \tag{8}$$

여기서 W_k와 F는 각각 핵심어 및 filler 모델을 의미하고 $O_{t_s}^{t_e}$는 프레임 t_s로부터 t_e까지의 관찰벡터 열을 나타낸다. 이 score를 적절한 문턱치와 비교하여 핵심어 여부를 최종적으로 판단하게 된다. 본 논문에서는 filler 모델에서 비핵심어 모델로 3 절에서 설명했던 GMM 및 monophone cluster를 각각 사용하여 실험을 하였다.

4.2. Anti-subword model을 이용한 방법

Subword 모델에 대한 anti-subword 모델[9]을 만들어 두면 인식대상 핵심어가 바

뀌는 상황에서도 anti-keyword를 자동으로 구성하여 후처리를 할 수 있다. 이 경우 특정 핵심어 모델 W_k가 N 개의 subword 모델로 구성될 경우 각각의 subword 모델에 대하여 미리 만들어 놓은 해당 anti-subword 모델을 사용하여 anti-keyword를 구성하게 된다.

본 논문에서는 훈련된 monophone 모델들 사이의 거리를 구해서 특정 monophone 모델에 대한 혼동 가능성이 높은 모델들을 이용하여 anti-subword 모델을 구성하였다. Anti-subword 모델을 구성하는 방법으로 2 가지 방식을 검토하였는데, 첫 번째 방식은 모델들 사이의 거리를 구할 때 분포간의 거리척도로써 weighted Euclidean distance를 사용하여 특정 모델에 대한 거리가 가장 작은 모델을 해당 anti-subword 모델로 정해준 방식이다. 이것을 수식으로 나타내면 다음과 같다.

$$\overline{p_i} = p_k$$

여기서

$$k = \arg\min_{j(j\neq i)} \left\{ \sum_{s=1}^{N} D_s(p_i, p_j) \right\} \tag{9}$$

이때 N은 모델의 상태수를 나타내고 V는 음성 특징벡터의 차원을 나타낸다. 상태의 분포간의 거리척도인 $D_s(p_i, p_j)$은 수식 (2)와 같다.

두 번째 방식은 첫 번째 방식과 동일하되 분포간의 거리척도로써 Kullback-Leibler distance를 사용한 방식이다. 이것을 수식으로 나타내면 다음과 같다.

$$\overline{p_i} = p_k$$

여기서

$$k = \underset{j(j\neq i)}{\operatorname{argmin}} \left\{ \sum_{s=1}^{N} \sum_{d=1}^{V} KL2\left(f_{isd}(x), f_{jsd}(x)\right) \right\} \tag{10}$$

이때 $KL2(f(x), g(x))$는 수식 (5)와 같고, N과 V는 모델의 상태 수와 음성 특징벡터의 차원을 나타낸다. 또한 $f_{isd}(x)$, $f_{jsd}(x)$는 각각 I 번째 및 j 번째 음소의 s 번째 상태에서의 d 번째 차원의 확률분포를 의미한다.

4.2.1. Anti-keyword score을 이용한 후처리 방법

Aniti-subword 모델을 사용하여 특정 핵심어 모델 W_k에 대한 anti-keyword를 구성하여 후처리를 하는 방법은 다음과 같다. 입력음성이 들어오면 전체 네트웍에서 핵심어를 찾게 되고, 이때 찾아진 핵심어 구간이 프레임 t_s로부터 프레임 t_e까지라고 하자. 인식된 핵심어의 log likelihood로부터 해당 anti-keyword 모델에 통과시켜 얻은 log likelihood를 뺀 값을 likelihood ratio score S_w라고 하면, 이는 다음 수식 (11)과 같이 표현된다.

$$S_{w_k} = \frac{1}{t_e - t_s}\left[\log P\left(O_{t_s}^{t_e} | W_k\right) - \log P\left(O_{t_s}^{t_e} | \overline{W}_k\right)\right] \tag{11}$$

여기서 $W_k (= p_1^{(k)} p_2^{(k)} \cdots p_N^{(k)})$ 는 keyword 모델이고 $\overline{W}_k (= \overline{p}_1^{(k)} \overline{p}_2^{(k)} \cdots \overline{p}_N^{(k)})$ 는 anti-keyword 모델이다. 이 score를 적절한 문턱치와 비교하여 인식결과를 수용 또는 거절하게 된다.

4.2.2. Anti-keyword의 subword segment를 이용한 후처리 방법

이 방법은 전체 네트웍을 통하여 검출된 핵심어 구간에 대하여 이 핵심어를 구성하는 subword의 각각의 구간을 해당 anti-subword에 다시 통과시켜서 확률값을 얻고, 이렇게 얻어진 확률값들을 더한 후 이 핵심어를 구성하는 subword의 개수로 나누어 준 것을 anti-keyword의 score로 사용한 방식이다. 인식결과가 $W_k = p_1^{(k)} p_2^{(k)} \cdots p_N^{(k)}$ 일 때 anti-keyword score $\log P\left(O_{t_s}^{t_e} | \overline{W}_k\right)$를 수식으로 나타내면 다음과 같다.

$$\log P\left(O_{t_s}^{t_e} | \overline{W}_k\right) = \frac{1}{N}\left[\log P\left(O_{t_0}^{t_1} | \overline{p}_1^{(k)}\right) + \log P\left(O_{t_1}^{t_2} | \overline{p}_2^{(k)}\right) + \cdots + \log P\left(O_{t_{N-1}}^{t_N} | \overline{p}_N^{(k)}\right)\right] \tag{12}$$

여기서 $O_{t_{N-1}}^{t_N}$ 은 핵심어를 구성하는 subword p_N 의 segment에 해당하는 특징벡터 열이다. 이렇게 해서 구한 anti-keyword 모델의 likelihood와 핵심어 모델의 likelihood의 차이를 수식 (12)와 같이 문턱치와 비교하여 인식결과를 수용 또는 거절하게 된다.

4.3. N-best decoding을 이용한 방법

이 방법에서는 N-best decoding을 이용하여 아래 식과 같이 best와 N-best 결과로부터 나온 best를 제외한 나머지 score들의 평균과의 차를 이용하게 된다[10].

$$S_{w_k} = \frac{1}{t_e - t_s}\left[\log P\left(O_{t_s}^{t_e} | W_k\right)_{1-best} - \frac{1}{N-1}\sum_{i=2}^{N} \log P\left(O_{t_s}^{t_e} | W_k\right)_{i-best}\right] \tag{13}$$

위 식에서 $1-best$ 와 $i-best$ 는 N-best decoding에 의해 얻어진 1 번째와 i 번째 hypothesis이고, $O_{t_s}^{t_e}$ 는 프레임 t_s 로부터 t_e 까지의 관찰벡터 열을 나타낸다. 여기서 제대로 인식된 경우는 1위와 나머지 후보간의 likelihood값의 차이가 크게 나며, 오인식된 경우는 차이가 상대적으로 적게 된다. 이 방법의 경우 적절한 N 값의 선정이 필요하며, 본 논문에서는 실험을 통해 N 값을 선택하였다.

4.4. Log likelihood를 이용한 신뢰도 정규화

지금까지 설명한 후처리 방식들은 식 (8), (11), (12) 및 (13)에서 보는 바와 같이

신뢰도, 즉, 귀무가설과 대립가설의 log likelihood ratio를 시간에 대해 정규화하여 사용하고 있으며, 이들의 일반적인 표현식은 다음과 같다.

$$\frac{1}{T}\left(\log p(\mathbf{O}|H_0)-\log p(\mathbf{O}|H_1)\right) \tag{14}$$

여기서 T는 핵심어 음성구간의 길이를, 그리고 H_0와 H_1은 핵심어 모델과 반모델을 각각 나타낸다.

Gupta 등은 시간에 대한 정규화 대신에 다음 식과 같이 신뢰도를 귀무가설에 해당하는 log likelihood 절대값의 크기로 정규화하는 방식을 사용하였으며, 이 방식이 시간에 대한 정규화 효과가 있으면서도 음성인식 과정에서의 문법 변화에 보다 잘 대처할 수 있다고 보고하였다[11].

$$\frac{\log p(\mathbf{O}|H_0)-\log p(\mathbf{O}|H_1)}{|\log p(\mathbf{O}|H_0)|} \tag{15}$$

본 논문에서도 개별적인 후처리 방식에 대해 시간에 대한 정규화 방식과 log likelihood 절대값 $|\log p(\mathbf{O}|H_0)|$로 정규화한 방식에 따른 실험을 각각 수행하여 그 성능을 비교하였다.

5. 실험 및 결과

5.1. 실험환경

가변어휘 핵심어 검출 실험은 다음과 같이 수행하였다. 음성특징 파리미터 추출은 음성신호를 16 kHz로 샘플링하여 20 msec 프레임 단위로 10 msec씩 shift하면서 전달함수가 $1-0.97z^{-1}$인 디지털 필터로 preemphasis를 하고, 여기에 다시 Hamming window를 씌운 후 12 차 Mel Frequency Cepstral Coefficient(MFCC) 및 delta 파라미터를 구하여 총 24 차 특징벡터를 사용하였다. Triphone 모델은 제한된 훈련 DB에 대해 모델 파라미터 추정의 신뢰도를 높일 수 있도록 tree based clustering을 사용하여 연속확률분포를 가지는 tied state HMM으로 훈련하였다. 모델당 상태수는 3 개를 사용하였으며 상태당 mixture 수를 변화시키면서 인식실험을 하였다. 음소모델의 훈련을 위해 한국전자통신연구원에서 구축한 음소열 최적화 단어 DB(POW 3848 DB)를 사용하였으며[11], 이 중에서 남자 40 명분을 사용하여 훈련하였다. 그리고 테스트를 위해, 한국전자통신연구소에서 구축한 부서명 DB 중 22 개의 유사성이 높은 부서명이 포함된 문장을 남성 15 명이 발성한 235 개 문장을 사용하였다.

5.2. 실험결과

5.2.1. 가변어휘 핵심어 검출 실험

비핵심어 모델링 방식에 다른 가변어휘 핵심어검출 실험결과를 표 1에 나타내었다. 표 1에서 비핵심어 모델로 GMM을 사용한 경우 mixture 수를 18 개에서 22 개까지 증가시켜 가면서 결과를 나타내었고, monophone clustering을 사용한 경우에도 역시 cluster 수를 가변시키면서 결과를 얻었다. 또한 핵심어 모델을 위한 tied state의 mixture 수는 1에서 8까지 증가시키면서 실험을 하였다.

표 1. 가변어휘 핵심어 검출 실험 결과

Filler model		Tied state mixture 수							
		1	2	3	4	5	6	7	8
Gaussian mixture 수	18	88.51	91.49	91.06	92.77	94.47	93.62	93.62	93.62
	20	88.51	91.49	91.91	94.04	93.62	**95.32**	93.62	93.19
	22	88.51	92.77	91.91	94.04	93.62	94.89	93.62	93.19
Monophone cluster 수 (Euclidean distance)	5	87.66	89.79	89.79	90.21	90.64	91.06	91.49	91.49
	7	86.81	88.09	90.21	**92.77**	92.77	92.34	91.91	92.77
	9	85.96	88.94	88.94	91.06	91.91	91.49	91.91	91.49
Monophone cluster 수 (Kullback distance)	7	86.81	88.51	89.36	91.49	91.91	91.49	91.91	91.49
	9	88.51	89.79	91.06	91.49	91.91	**92.34**	91.91	91.49
	11	88.94	89.79	90.64	91.49	91.91	91.49	91.91	91.91

실험 결과를 보면 monophone clustering 방식의 경우 weighted Euclidean distance와 Kullback-Leibler distance를 사용한 결과의 성능 차이가 별로 없었으며, 비핵심어로 GMM을 사용한 방식이 monophone clustering을 사용한 방법보다 전반적으로 우수한 성능을 보였다. GMM을 filler 모델로 사용하는 경우 Gaussian 수가 20 개인 경우 95.32%로 가장 좋은 성능을 나타내었다.

5.2.2. 후처리 방식에 따른 실험 및 결과

후처리 실험은 앞서 설명한 비핵심어 모델 실험에서 가장 좋은 성능을 얻은 경우에 대해 수행하였다. 즉, 가변어휘 핵심어 검출에서 비핵심어 모델로는 20 개의 mixture를 가지는 GMM을 사용하고, tied state의 mixture 수를 6 개로 사용하는 경우에 대하여 실험을 하였다. 시간으로 정규화한 여러 가지 후처리 방법에 의해 구해진 핵심어 기각률에 대한 핵심어 인식성능을 그림 4에 나타내었다.

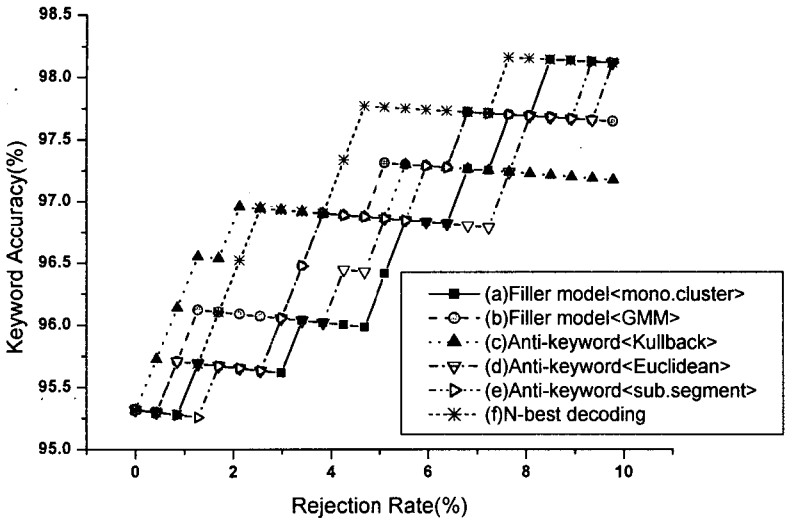

그림 4. 시간으로 정규화한 후처리 방식들의 성능 비교

그림 4에서 (a)는 filler 모델을 사용하는 후처리 방식에서 filler 모델로 9 개의 cluster를 가지는 monophone cluster를 사용한 경우이며, (b)는 20 개의 mixture를 가지는 GMM을 사용한 경우이다. (c)는 anti-keyword 모델을 사용하는 후처리 방식에서 anti-subword를 구성할 때 Kullback-Leibler 거리를 사용하여 혼동 가능성이 가장 높은 모델 하나를 해당 anti-subword로 사용한 경우이고 (d)는 (c)와 나머지는 동일하되 anti-subword를 구성하기 위한 모델들의 확률분포 사이의 거리척도로써 weighted Euclidean distance를 사용한 것이다. (e)는 (c)의 방법으로 anti-keyword를 구성하고 여기에 subword segment별 score를 이용하여 후처리를 한 방식이다. 마지막으로 (f)는 N-best decoding을 결과를 이용한 후처리 방식으로서, 본 실험에서는 N를 1에서 10까지 변화시켜 가면서 실험을 한 결과 가장 좋은 성능을 보이는 3-best 결과를 나타내었다.

Anti-keyword 모델 기반의 방식으로 Kullback-Leibler distance를 통해 anti-subword로 사용한 경우의 성능이 핵심어 기각률이 2% 이하로 작은 경우에 대해 가장 우수하였다. Filler 모델 기반의 방식에서는 monophone clustering를 사용하는 방식보다는 GMM을 사용하는 방식이 좀 더 나은 성능을 나타냈으며, 이는 앞서 비핵심어 모델링 실험의 결과와도 일맥상통하는 부분이다. N-best decoding의 경우 2% 이상의 기각률에서 가장 우수한 성능을 나타내었다.

그림 5에는 4.4 절에서 언급한 바와 같이 시간 정규화 대신 log likelihood의 절대값으로 정규화한 후처리 방식들의 성능을 나타내었다.

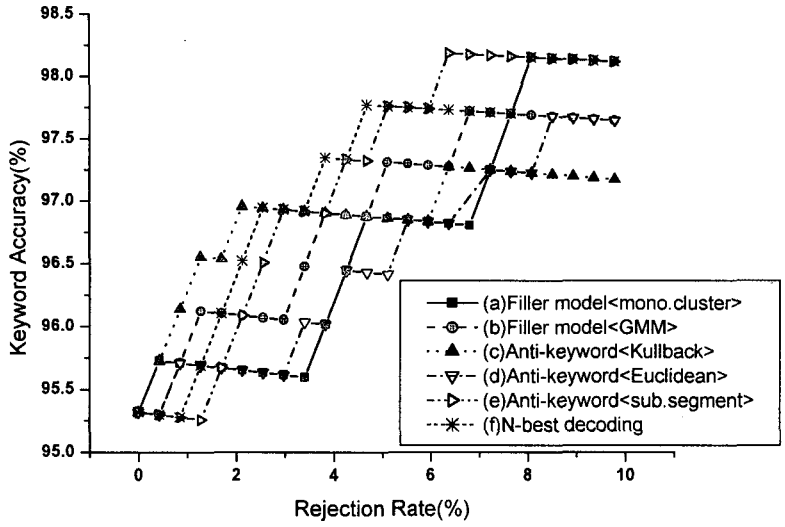

그림 5. Log likelihood로 정규화한 후처리 방식들의 성능 비교

그림 4와 그림 5를 비교해 볼 때 시간 정규화 및 log likelihood에 의한 정규화의 성능은 대체로 비슷하였으며, GMM을 이용한 filler 모델과 anti-keyword의 subword segment에 따른 후처리 방식에서 시간 정규화 방식보다 log likelihood에 의한 정규화 방식이 더 나은 성능을 보였다. 그리고, 5% 이상의 핵심어 기각률이 실제 응용에 적용하기 곤란함을 고려할 때, 2% 이내의 핵심어 기각률에서는 Kullback-Leibler distance에 의한 anti-subword 모델을 이용한 방법이, 그리고 대략 2%에서 5% 사이에서는 N-best decoding에 의한 방식이 가장 우수하다는 점에서는 두 가지 정규화 방식이 일치된 결과를 나타내었다. 핵심어 기각률을 5% 허용할 경우 가장 우수한 성능을 보인 N-best decoding을 이용한 방식을 사용하여 50% 이상의 오류율 감소를 얻을 수 있었다.

6. 결 론

본 논문에서는 가변어휘 핵심어 검출기를 구현하고, 몇 가지 비핵심어 모델링 방법 및 후처리 방법의 성능을 평가하였다. Monophone clustering 및 GMM을 이용한 두 가지 비핵심어 모델링 방법을 검토한 결과, GMM을 사용한 방법이 monophone clustering 방법보다 더 우수한 성능을 보였다. GMM 비핵심어 모델을 사용한 경우 95.32%의 핵심어 인식성능을 얻을 수 있었다. 후처리 과정에서 anti-subword 모델을 구성함으로써 가변어휘 상황에서도 anti-keyword를 구성할 수 있도록 하였으며, likelihood ratio 형태인 신뢰도의 정규화 방식으로 시간 정규화와 likelihood에 의한 정규화를 적용하였다. 실험 결과 기각률 2% 이내의 경우 Kullback-Leibler 거리를 이용한 anti-subword 모델링 방

식이 가장 우수하였고, 2%에서 5% 사이의 기각률에서는 N-best decoding 방법의 성능이 우수하였다. 핵심어 기각률이 5%일 때 N-best decoding 방식이 50% 이상의 오류율 감소를 보여 filler 모델 및 anti-subword 모델 방식보다 더 나은 성능을 나타냈다.

참 고 문 헌

[1] 김형순. 1994. "Keyword spotting 기술." *한국통신학회지*, 11(9), 57-64.

[2] Wilpon, J. G., L. R. Rabiner, C. H. Lee & E. R. Goldman. 1990. "Automatic recognition of keywords in unconstrained speech using hidden Markov models." *IEEE Trans. Acoust., Speech, Signal Processing*, 38(11), 1870-1878.

[3] Rose, R. C. & D. B. Paul. 1990. "A hidden Markov model based keyword recognition system." *Proc. IEEE ICASSP*, 129-132.

[4] 이활림, 김형순, 김유신. 1997. "음소 HMM을 이용한 핵심어 검출 시스템의 성능향상에 관한 연구." *한국음향학회지*, 16(8), 60-67.

[5] Wilpon, J. G. & L. R. Rabiner. 1985. "A modified K-Means clustering algorithm for use in isolated word recognition." *IEEE Trans. Acoust., Speech, Signal Processing*, 33(3), 587-594.

[6] Li, X. Q. & I. King. 1999. "Gaussian mixture distance for information retrieval." *IJCNN*, 2544-2549.

[7] Kullback, S. & R. Leibler. 1951. "On the information and sufficiency." *Annals of Mathematical Statistics*, 22, 79-86.

[8] Reynolds, D. A. & R. C. Rose. 1995. "Robust text-independent speaker identification using Gaussian mixture speaker models." *IEEE Trans. Speech and Audio Processing*, 3(1), 72-83.

[9] Sukkar, R. A. & C. H. Lee. 1996. "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition." *IEEE Trans. Speech and Audio Processing*, 4(6), 420-429.

[10] Setlur, A. R., R. A. Sukkar & J. Jacob. 1996. "Correcting recognition errors via discriminative utterance verification." *Proc. ICSLP*, 602-605.

[11] Gupta, S. K. & F. K. Soong. 1998. "Improved utterance rejection using length dependent thresholds." *Proc. ICSLP*, 795-798.

[12] Lim, Y. J. & Y. J. Lee. 1995. "Implementation of the POW (phonetically optimized words) Algorithm for Speech Database." *Proc. IEEE ICASSP*, 89-92.

접수일자: 2003. 7. 28.
게재결정: 2003. 8. 30.

▲ 김형순
부산시 금정구 장전동 산 30번지 (우: 609-735)
부산대학교 전자공학과 음성통신실험실
Tel: +82-51-510-2452 Fax: +82-51-515-5190
E-mail: kimhs@pusan.ac.kr

▲ 김영국
부산시 금정구 장전동 산 30번지 (우: 609-735)
부산대학교 전자공학과 음성통신실험실
Tel: +82-51-510-1704 Fax: +82-51-515-5190
E-mail: ykukim@pusan.ac.kr

▲ 신영욱
부산시 금정구 장전동 산 30번지 (우: 609-735)
부산대학교 전자공학과 음성통신실험실
Tel: +82-51-510-1704 Fax: +82-51-515-5190
E-mail: space73@pusan.ac.kr