다이폰 기반의 Generic Word Model을 이용한 거절 알고리즘

A Study on the Rejection Algorithm Using Generic Word Model Based on Diphone Subword Unit

정 익 주* · 정 훈**
Ikjoo Chung · Hoon Chung

ABSTRACT

In this paper, we propose an algorithm on OOV (Out-of-Vocabulary) rejection based on two-stage method. In the first stage, the algorithm rejects OOVs using generic word model, and then in the second stage, for further reduction of false acceptance, it rejects words which have low similarity to the candidate by measuring the distance between HMM models. For the experiment, we choose 20 in-vocabulary words out of PBW445 DB distributed by ETRI. In case that the first stage is processed only, the false acceptance is 3% with 100% correct acceptance, and in case both stages are processed, the false acceptance is reduced to 1% with 100% correct acceptance.

Keywords: Generic Word Model, Out-of-Vocabulary Rejection

1. 서 론

음성은 인간의 가장 자연스러운 통신 수단으로 맨-머신(man-machine) 인터페이스에서 중요한 역할을 담당할 것으로 기대되어 왔지만 현재의 상황은 그렇지 못한 실정이다. 이렇게 음성인식 기술이 실용적으로 사용되지 못하는 데는 아직 기술적으로 해결해야 할 문제가 많음을 의미한다. 이중 거절기능은 실용적인 음성인식 시스템이 되기 위해서는 반드시 필요한 기능으로 인식 대상 어휘(In-Vocabulary)가 아닌 음성이 입력된 경우 이를 인식 대상 어휘가 아니라고 판별함으로써 음성 인식 시스템이 오동작하는 것을 막아주는 역할을 한다. OOV의 처리는 크게 거절을 위한 별도의 모델을 사용하는 방식과 적절한 후처리 과정을 통한 방식으로 구분된다. 별도의 모델을 사용하는 방식에는 필러(Filler)모델[4]과 반단어(Anti-Word)모델을 사용하는 방식이 있다. 필러 모델은 핵심어 방식[1]의 음성 인식기에서 주로 사용하는 방식으로 인식 대상이 아닌 단어나 묵음을 흡수하기 위해 문맥 독립형 서브 워드 모델을 연결하거나 음성학적으로 의미있는 모델들을 타잉(tying)하여 필러 모델을 구성한다. 반단어(Anti-Word) [2,3]모델은 인식대상 어휘에 반대가 되는 모델을 구성하여 인식 대상에 포함시키는 방식으로 False Acceptance가 발생할 가능성이 높은 OOV를 인식 대상 어휘와 동일한

* 강원대학교 전기전자정보통신공학부
** 강원대학교 전자공학과

도메인 상에서 인식함으로써 OOV를 제거하는 방식이다. 반단어 모델은 태스크 도메인이 결정되어 인식 대상 어휘가 고정되어 있는 경우에 주로 사용되는데 태스크 도메인 상에서 발생할 수 있는 OOV에 대한 별도의 단어 모델을 구성하여 OOV를 제거하게 된다. 후처리 과정을 사용하는 방법에는 likelihood ratio에 의한 방식과 confidence scoring과 같은 방식이 있다. likelihood ratio 방식에는 인식된 결과가 일정 문턱값 이하로 떨어지는 경우에는 OOV로 간주하는 방식으로, 서브 워드 모델의 모델링 정도에 따라 성능이 많은 영향을 받게 된다. 더욱이 특정 서브 워드 모델은 훈련 DB가 풍부하여 모델링이 잘되고 특정 모델은 그렇지 못하다면 거절을 위한 문턱값을 설정하는데 어려움이 발생한다. confidence scoring 방식에서는 인식된 결과가 OOV인지 인식 대상 어휘인지를 인식시에 사용한 파라메타 값들의 스코어값에 근거하여 판별하는 방식이다. 본 논문에서는 2 단계에 걸친 거절 알고리즘에 대해 설명한다. 첫 번째 방식은 다이폰 기반의 Generic Word Model[6]이고 두 번째 방식은 HMM (Hidden Markov Model) 간의 유사도를 사용한 Confidence Scoring 방식이다. OOV의 처리는 인식 대상 어휘가 유한개인 경우에 필요하다. 만일 인식 대상 어휘의 범위가 가변 어휘로 표현 가능한 모든 단어라면 OOV란 개념은 없어지며 단지 현재 인식 하고자 하는 어휘와 그렇지 않은 어휘가 인식 대상 어휘와 OOV를 대신하게 된다. 이런 인식 시스템에서는 얼마나 정확하게 인식하느냐와 같은 인식 시스템의 성능이 결국은 거절 기능의 성능 역시 결정하게 된다. 본 논문에서 제안하는 첫 번째 방식의 거절 알고리즘에서는 Generic Word Model이 가변 어휘 방식에서 표현 가능한 모든 단어를 표현하도록 구성함으로써 거절 기능을 단순히 대규모 어휘의 음성 인식 시스템의 입장에서 해석하도록 한다. 첫 번째 단계에서는 거절 알고리즘의 성격상 False Acceptance가 발생할 수 있으므로 두 번째 단계에서는 첫 번째 단계를 통해 얻어지는 2 개의 HMM열에 대해 HMM간 유사도에 기반을 둔 confidence scoring[5,6]을 통해서 거절기능을 한번 더 수행하게 된다. 2 장에서는 Generic Word Model의 개념과 기존에 구현된 내용에 대해 살펴본다. 3, 4 장에서는 본 연구에서 제안한 서브 워드 모델 기반의 Generic Word Model과 Confidence Scoring 알고리즘에 대해 살펴본다. 5장에서는 본 논문에서 제안한 알고리즘의 성능을 평가하기 위해 다이폰을 사용한 Generic Word Model을 사용한 경우와 다이폰을 기반으로 하는 필러 모델을 사용한 Generic Word Model의 사용한 경우에 대한 거절 성능에 대한 실험을 수행하였다.

2. Generic Word Model을 사용한 거절 기능

거절기능이란 음성 인식 시스템에서 인식 대상 어휘(In-Vocabulary)가 아닌 단어가 입력된 경우에 이를 인식 대상 어휘가 아님을 알려주는 기능을 의미한다. 이때 별도의 모델을 기반으로 하는 거절 기능에서는 필러나 반단어 모델을 사용하게 되는데 이렇게 OOV를 표현하기 위한 모델을 Generic Word Model이라고 한다. Generic Word Model은 그림 1과 같이 서브 워드 모델을 연결 단어 형태로 구성하거나 그림 2와 같이 인식 대상 어휘에 반대가 되는 반단어(Anti-word)를 등록하여 구성할 수도 있다. 가변 어휘 방식의 인식기와 같이 인식 대상 어휘가 고정되어 있지 않는 경우에는 그림 1과 같은 형태의 Generic Word Model을 주로

사용하고 인식 대상 어휘가 고정되어 있는 경우에는 그림 2와 같은 반단어 모델이 널리 사용되고 있다.

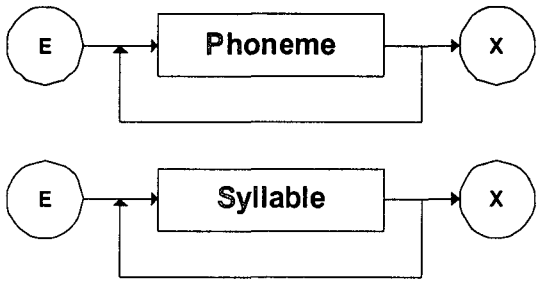

그림 1. 연결 단어 방식의 Generic Word Model

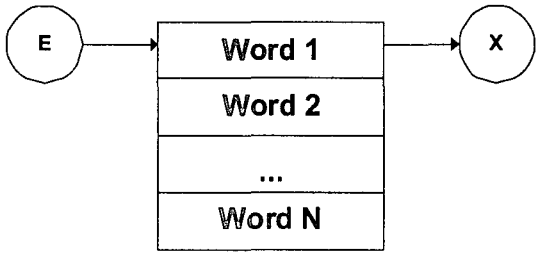

그림 2. 반단어(Anti-Word) 방식의 Generic Word Model

그림 1과 같이 서브워드 모델을 연결하는 경우에는 주로 문맥 독립적인 음소 모델을 사용하며 반단어 모델에서는 문맥종속적인 모델을 사용하여 Generic Word Model을 구성하게 된다. 거절 기능의 성능은 결국은 인식 성능에 의해 결정되므로 만일 반단어에 대한 충분한 정보가 제공된다면 문맥종속적인 서브 워드 모델을 사용하는 반단어 기반의 Generic Word Model이 문맥 독립적인 서브 워드 모델을 사용하는 방식에 비해 좀더 나은 거절 성능을 보일 것이다. 최근 널리 사용되고 있는 문맥 종속 모델로는 트라이폰이 있으나 모델의 개수가 만개 이상이 되므로 이를 이용하여 그림 1과 같은 형태의 Generic Word Model을 구성하는 것은 현실적으로 불가능하다. 따라서 본 연구에서는 모델의 정밀도와 개수면에서 문맥 독립 음소 모델과 트라이폰 모델의 중간쯤에 위치하는 다이폰을 사용하여 그림 1과 같은 Generic Word Model을 구성한다. 이때 다이폰이 반복적으로 인식될 수 있도록 구성하여 가변어휘에서 표현 가능한 모든 단어를 인식 대상으로 하는 Generic Word Model을 구성함으로써 궁극적으로는 그림 2에서 N이 무한대인 반단어 모델을 구성한다.

3. 다이폰 기반의 Generic Word Model

다이폰이란 한 음소의 안정 구간에서 다음 음소의 안정 구간 사이에서 발생하는 음성 신호의 변이 현상을 모델링한 인식 단위로 한국어의 경우 대략 1,400 개 정도가 존재한다.

3.1 다이폰의 형태

다이폰은 크게 CV(자음+모음), VC(모음+자음), CC(자음+자음), VV(모음+모음) 형태가 존재하며 이들 모델의 조합으로 단어 모델이 구성되는데 그림 3은 다이폰을 결합하여 단어 모델을 구 성하는 과정을 도시적으로 보여주고 있다.

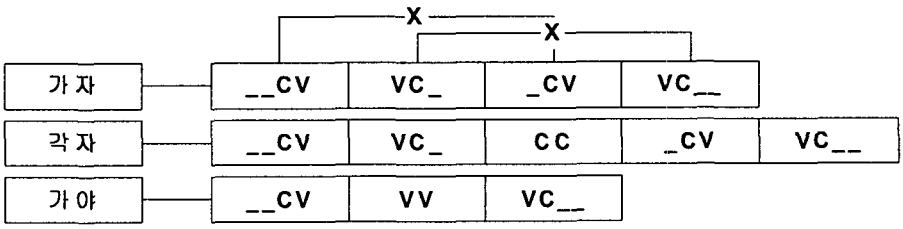

그림 3. 다이폰 형태들

본 연구에서는 그림 3에서 같이 CV형과 VC형은 다이폰이 놓이는 위치에 따라 다르게 구분되는데 __CV형은 묵음+CV 형태의 다이폰을 모델링하고 _CV형은 음소+CV 형태의 다이폰을 모델링한다. VC형의 경우도 VC__은 VC+묵음을 모델링하고 VC_는 VC+음소의 다이폰을 모델링한다.

3.2 다이폰을 사용한 Generic Word Model 구성

가변 어휘 방식에서 Generic Word Model은 인식 대상 어휘가 아닌 모든 형태의 단어를 모델링하기 위한 일반적인 구조를 지녀야 하므로 그림 4와 같이 1 음절에서 N 음절로 구성된 단어에 대한 Generic Word Model을 구성할 수 있다.

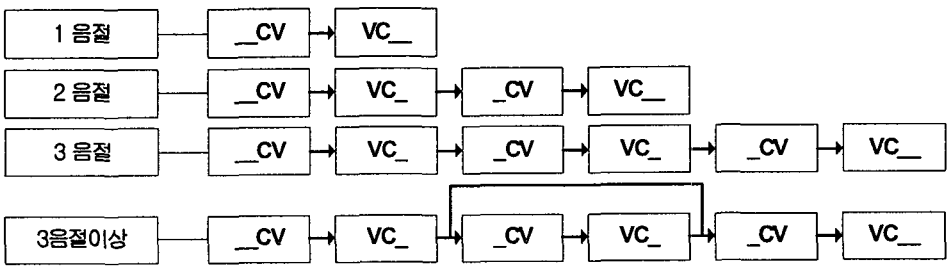

그림 4. 다이폰을 이용한 Generic Word Model 구성

그림 4를 기반으로 본 연구에서 사용한 인식 대상 어휘(In-Vocabulary)와 OOV(Out-of-

Vocabulary)를 동시에 포함한 인식 네트워크는 그림 5와 같다.

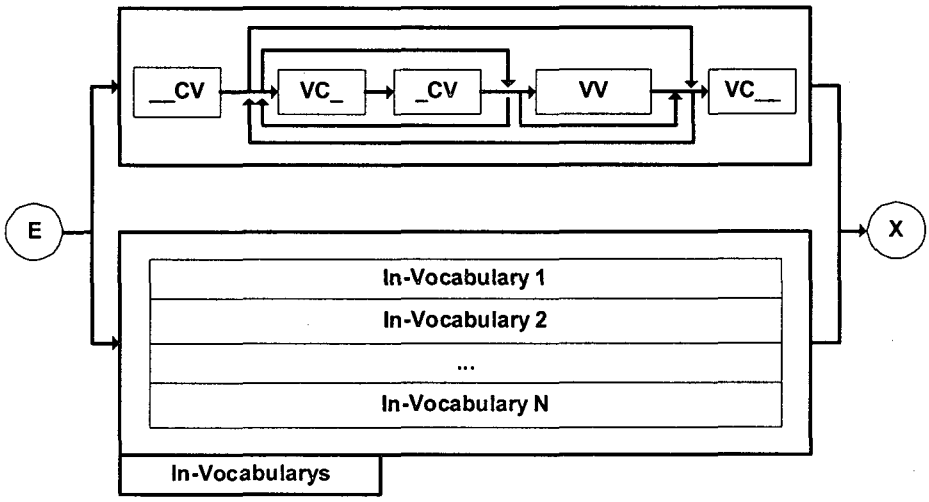

그림 5. 거절 기능을 포함한 인식 네트워크

3.3 거절 알고리즘

필러 모델을 사용하는 일반적인 거절 알고리즘에서는 필러 모델이 인식 대상 어휘를 포함하지 않도록 설계하여 인식 대상 어휘가 OOV로 빠지는 것을 방지하고 식 (1)과 같이 인식 대상 어휘에 대한 관측열의 log-likelihood값이 필러 모델(여기서는 Generic Word Model)에 대한 관측열의 log-likelihood값보다 일정 문턱값 이상 되는 경우를 인식 어휘로 간주한다.

$$\log(P(O \mid W_{in-vocabulary})) - \log(P(O \mid W_{GenericWordModel})) \geq threshold \quad (1)$$

그러나 이런 방식에서는 필러의 선택이 인식 대상 어휘에 의존적이므로 인식 대상 어휘가 변하게 되면 필러 역시 변경되어야 한다. 그러나, 본 연구에서는 모든 표현 가능한 단어를 모델링할 수 있는 Generic Word Model을 사용하므로 인식 어휘가 변경되더라도 OOV에 대한 모델을 변경할 필요가 없다.

$$abs(\log(P(O \mid W_{in-vocabulary})) - \log(P(O \mid W_{GenericWordModel}))) \leq threshold \quad (2)$$

단, 거절 알고리즘이 식 (2)과 같이 변경되어 사용된다. 인식 대상 어휘를 표현하기 위해 사용하는 서브 워드 모델인 다이폰을 사용해 Generic Word Model을 구성하므로 인식 대상 어휘는 반드시 Generic Word Model 내에 포함되게 된다. 따라서, 인식 대상 어휘가 발음된 경우에는 인식된 두 log-likelihood 값의 차이는 적어지며 인식 대상 어휘가 발음되지 않은 경우에 두 log-likelihood값의 차이는 커지게 된다. 따라서, 식(2)와 같이 값의 차이가 일정 문턱

값 내에 존재하면 발음된 음성을 인식 대상 어휘로 간주하게 된다. 이때 차이에 절대값을 취한 것은 인식기가 100% 정확하지 않으므로 인식 대상 어휘가 발음되더라도 인식된 결과값이 Generic Word Model의 값보다 크다고 보장하지 못하고 그 차이가 얼마인지가 거절에서 의미를 지니게 된다. 그러나, 2 단계 Confidence Scoring에서는 식(1)의 알고리즘을 사용하여 최종 거절 기능을 수행하게 된다.

4. HMM 간의 유사도 측정을 통한 Confidence Scoring

그림 5의 인식 네트워크를 사용하면 입력되는 음성 신호에 대해 그림 6과 같이 디코딩된 2 개의 서브 워드 열에 대한 정보를 얻을 수 있다. 하나는 인식된 어휘로부터 얻을 수 있으며 또 하나는 Generic Word Model의 디코딩 결과로 얻을 수 있다. 입력된 음성이 인식 대상 어휘이고 서브 워드 모델이 정확히 훈련되었다면 디코딩된 2 개의 서브워드 열은 동일하거나 유사한 형태가 될 것이며 본 절에서는 이 유사도를 측정하여 OOV를 거절하게 된다.

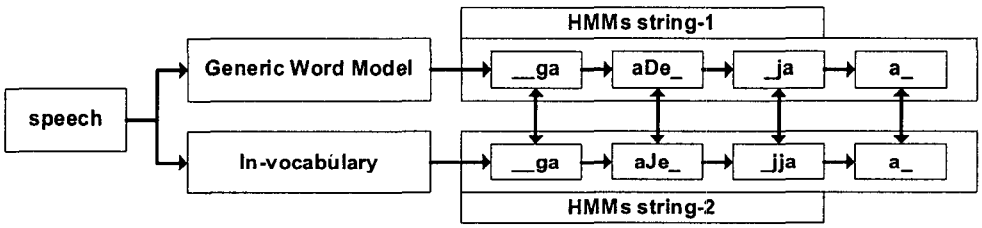

그림 6. 입력된 음성에 대한 디코딩된 2 개의 HMM열

2 개의 HMM 모델열 간의 유사도를 측정하기 위해서는 HMM 간의 유사도를 측정하는 함수와 열의 개수가 틀릴 수 있으므로 길이를 보상하기 위한 기능이 필요하다. HMM 간의 유사도는 식 (3)과 같이 HMM의 평균 벡터를 관측열이라 간주하고 각각 모델에 대한 관측 확률값을 구해 평균을 취했다.

$$Dist(ix,iy) = \frac{1}{2}\sum_{j=1}^{N}(b_{ix,j}(iy_m) + b_{iy,j}(ix_m)) \tag{3}$$

$$D_A(ix,iy) = \min_{(ix',iy')}(D_A(ix,iy) + Dist(ix,iy)) \tag{4}$$

길이를 보상하기 위해서는 식(4)와 같이 DTW를 사용하였다. 그림 7은 2 개의 HMM열을 DTW를 사용해 유사도를 측정하는 과정을 도식적으로 설명한다.

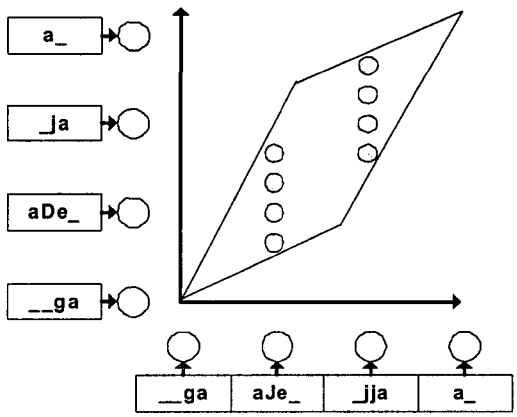

그림 7. DTW를 이용한 HMM열 간의 유사도 측정

5. 인식 실험 및 결과

본 실험에서는 3 장에서 제안한 기본 인식 모델을 사용하여 Generic Word Model을 구성한 경우의 거절 성능이 일반적인 필러(Filler)모델을 사용하여 Generic Word Model을 구성한 것에 비해 어느 정도의 성능 향상이 이루어졌는지에 대한 평가와 후처리 과정으로 HMM 모델간의 Confidence Scoring을 적용한 경우에 거절 기능의 향상에 대해 평가해 보기 위함이다. 다음은 거절 성능 평가를 위한 실험 환경이다.

1. 음성 신호 해상도: 16 KHz로 샘플링되었으며 샘플당 16 bit의 해상도를 지닌다.
2. 특징 벡터 분석: 25 ms의 분석 구간에 12.5 ms로 오버랩핑하였다. 특징 벡터는 12 차 MFCC에 C0에너지를 사용하였으며 델타 파라메타를 사용해 총 26 차로 구성되었다.
3. 기본 인식 모델: 1,400 개의 다이폰 모델을 사용하였으며 ETRI에서 배포한 PBW445DB, POW3848와 국어 공학 연구소의 PBW DB를 사용하여 훈련되었다.
4. 필러-다이폰: 다이폰 모델을 타잉(tying)하여 145 개의 필러-다이폰을 생성하였다. 다음과 같은 음소를 타잉(tying)하였다.

   초성: (ㄱ, ㄲ, ㅋ), (ㄴ), (ㄷ,ㄸ,ㅌ), (ㄹ), (ㅁ), (ㅂ, ㅃ, ㅍ), (ㅅ, ㅆ), (ㅇ), (ㅈ, ㅉ, ㅊ), (ㅎ)
   중성: 타잉 없음.
   종성: (ㄱ, ㄷ, ㅂ), (ㄴ, ㄹ, ㅁ, ㅇ)

5. 테스트 환경: ETRI의 PBW445DB 중 ASW 화자가 발음한 2 세트의 445DB를 사용하였다. 445 단어 중 다음과 같이 임의로 선정한 20 개의 단어를 인식 대상 어휘(In-Vocabulary)로 선정하였으며 나머지 425 개의 단어를 OOV로 간주하였다. 인식 대상은 주로 2, 3 음절을 위주로 선정하였으며 선정된 단어는 다음과 같다.

가운데, 고유하다, 교육, 금붕어, 꽃송이, 내륙, 느낌, 도읍지, 둘러앉다, 뜨개, 매우, 물벼룩, 발자취, 볏단, 비용, 뻘뻘, 쓰임새, 아파트, 약수터, 옛날

5.1 실험 결과

3 절과 4 절에서 제안한 알고리즘에 대한 거절 성능의 결과는 다음과 같다.

5.1.1 다이폰 기반의 Generic Word Model을 적용한 실험 결과

그림 8과 9에서는 문턱값의 변화에 따른 필러-다이폰을 사용한 Generic Word Model과 다이폰 모델을 사용하여 Generic Word Model을 구성한 경우의 거절 성능을 보여준다.

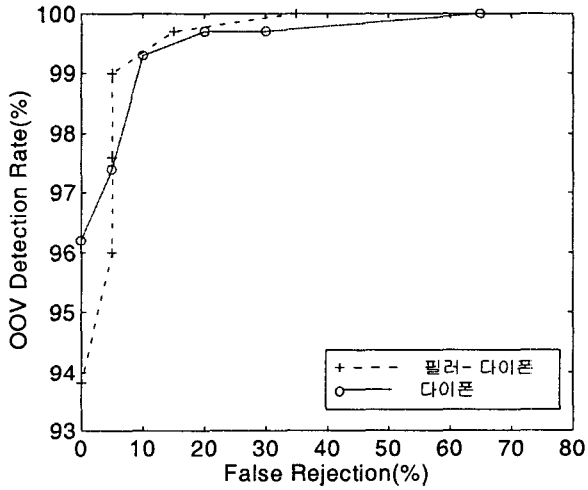

그림 8. 문턱값 변화에 따른 OOV Detection과 False Rejection 값의 변화

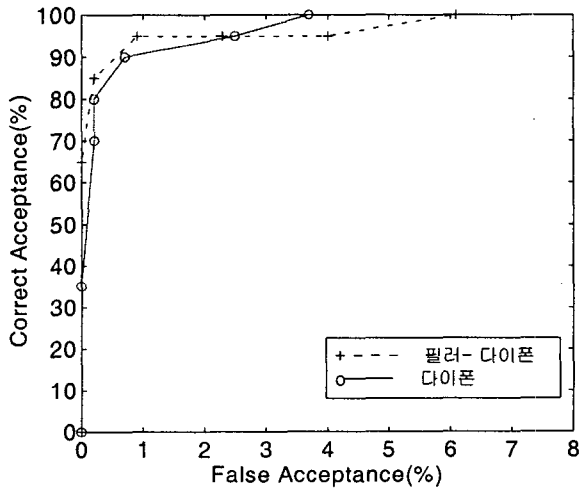

그림 9. 문턱값 변화에 따른 Correct Acceptance와 False Acceptance 값의 변화

위의 그래프에서 보듯이, 첫 번째 단계만을 적용하였을 경우 기존의 방법(필러를 이용하는 방법)과 대동 소이한 성능을 보인다. 한편, 다이폰을 사용한 Generic Word Model의 경우에는 100%의 Correct Acceptance를 얻기 위해서는 3.8%의 False Acceptance가 발생되었으며 필러-다이폰을 사용한 경우에는 6% 가량의 False Acceptance가 발생하였다. 따라서 높은 Correct Acceptance가 요구되어지는 응용에서는 다이폰을 사용한 Generic Word Model이 유리하다.

5.1.2 모델간의 유사도에 기반한 Confidence Scoring을 적용한 실험 결과

5.1.1에서 발생하는 False Acceptance는 식 (2)와 같이 문턱값을 사용하여 거절 기능을 수행하는 경우에는 불가피하게 발생하게 된다. 따라서, 4 장에서 제안한 알고리즘을 사용하여 두 번째 단계인 인식된 단어의 검증 과정을 거치게 된다. 그림 10과 그림 11은 4 장에서 제안한 HMM 모델간의 유사도에 기반한 Confidence Scoring 방식을 적용한 결과이다. 두 그래프에서 보듯이 Confidence Scoring 단계를 거친 경우 첫 번째 단계만을 거친 경우 보다 좋은 성능을 보일 뿐만 아니라, 필러를 이용하는 기존의 방식보다 우수한 성능을 보인다. 뿐만 아니라, 첫 번째 단계만 거쳤을 때와 비교하면, Correct Acceptance는 기존의 100%로 유지하면서도 False Acceptance가 3.8%에서 1.2%로 줄어들었으며 False Rejection도 65%에서 20%로 줄어들었다.

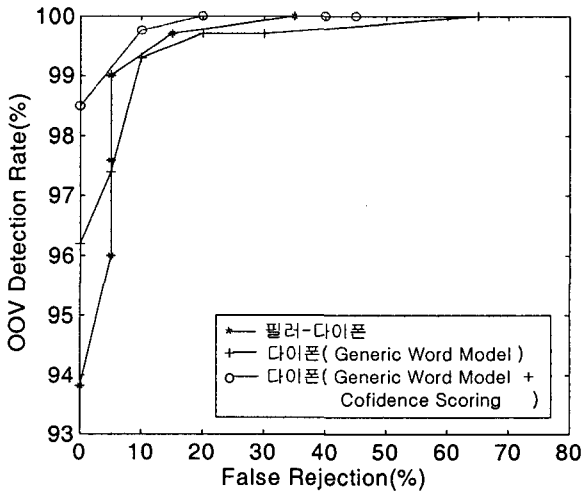

그림 10. 문턱값 변화에 따른 OOV Detection과 False Rejection 값의 변화

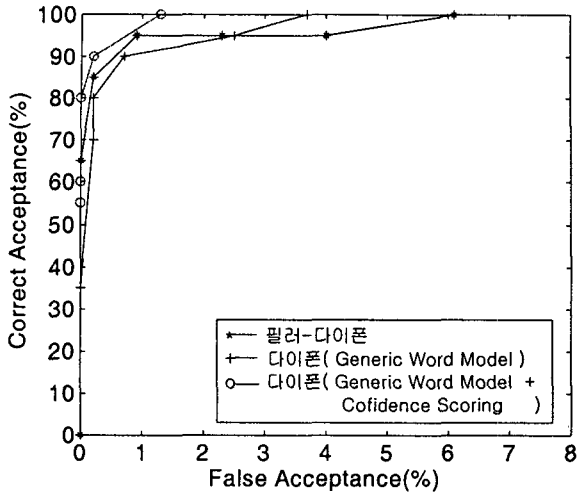

그림 11. 문턱값 변화에 따른 Correct Acceptance와 False Acceptance 값의 변화

6. 결 론

본 연구에서는 2 단계의 과정을 거쳐 거절 기능을 구현하였다. 첫 번째 단계에서는 다이폰 모델을 연결한 Generic Word Model을 사용하였다. 문맥 종속적인 모델을 사용하여 단어 모델을 표현하고 문맥 독립적인 모델을 사용하여 Generic Word Model을 구성하는 일반적인 거절 알고리즘과는 달리 단어 모델을 구성하는데 사용한 서브 워드 모델을 사용하여 Generic Word Model을 구성하였다. 이는 Generic Word Model이 의미하는 바와 같이 서브 워드 모델로 표현 가능한 모든 단어에 대한 모델링이 가능하게 함으로써 거절 기능을 단지 대규모 어휘의 음성 인식기의 입장에서 현재 인식 가능한 어휘와 그렇지 않은 어휘로 구분하는 것과 동일한 효과를 가지게 하였다. 따라서, 거절 기능을 전적으로 인식기의 성능에 종속적으로 만듦으로써 거절 기능을 구현하기 위한 별도의 작업이 필요치 않으면서도 가장 이상적인 형태의 Generic Word Model을 구성하도록 하였다. 첫 번째 거절 알고리즘의 특성상 False Acceptance는 불가피하게 발생하므로 두 번째 단계에서는 입력된 신호에 대해 Generic Word Model을 통해 디코딩된 서브 워드 열과 인식 대상 어휘를 통해 디코딩된 서브 워드 모델열 간의 유사도를 측정하여 첫 번째 단계에서 인식된 단어에 대해 추가적인 검증과정을 거쳤다. 결과적으로 이 두 단계의 알고리즘을 사용함으로써 Correct Acceptance는 최대치를 유지하면서 False Acceptance와 False Rejection을 효과적으로 줄일 수 있었다. 본 연구에서 제안한 방식의 Generic Word Model에서는 인식 성능의 향상과 거절 성능의 향상은 밀접한 관계를 가지게 되므로 보다 문맥 종속적이 모델로 사용한다면 거절 기능이 향상될 것으로 여겨진다. 아울러 인식 성능을 향상시키는 알고리즘을 같이 사용한다면 좀더 나은 성능의 거절 성능을 기대할 수 있다.

참 고 문 헌

[1] 김형순. 1994. "Keyword Spotting 기술." *한국통신학회지*, 제11권 9호, 57-64.
[2] 김동화, 김형순, 김영호. 1997. "고립단어인식 시스템에서의 거절기능 구현." *한국음향학회지*, 16권 6호, 106-104.
[3] 김동화, 김형순, 김영호. 1997. "집단화된 음소 모델을 이용한 거절기능 구현." *제10회 신호처리합동학술대회 논문집*, 제10권 1호, 701-704.
[4] 김우성, 구명완. 1999. "반음소 모델링을 이용한 거절기능에 대한 연구." *한국음향학회지*, 제18권 3호.
[5] Kamppari, S. & T. Hazen. 2000. "Word and phone level acoustic confidence scoring." *Proc. of ICASSP.*
[6] Bazzi, Issam. 2002. *Modelling Out-of-Vocabulary Words for Robust Speech Recognition.* MIT. Doctoral Thesis.

접수일자: 2003. 3. 28.
게재결정: 2003. 5. 13.

▲ 정익주
강원도 춘천시 효자2동 (우: 200-701)
강원대학교 전기전자정보통신 공학부
Tel: +82-33-250-6322
E-mail: ijchung@kangwon.ac.kr

▲ 정훈
강원도 춘천시 효자2동 (우: 200-701)
강원대학교 전자공학과
Tel: +82-33-250-6322
E-mail: hchung@kwnu.kangwon.ac.k