반음소 모델링을 이용한 거절기능에 대한 연구

A Study on the Rejection Capability Based on Anti-phone Modeling

김 우 성*, 구 명 완*
(Woo Sung Kim*, Myoung Wan Koo*)

요 약

본 논문에서는 독립단어 음성인식 시스템을 위하여 반음소(anti-phone) 모델링을 이용한 인식 거절(rejection)기능에 대해 기술한다. 음성인식 거절 기능은 음성인식기를 제작할 때 정해놓은 인식대상 단어 이외의 단어가 입력되었을 때 그 단어가 인식할 수 없는 단어임을 알려주는 기능이다. 음성인식 거절기능을 구하는 방식은 핵심어 검출(keyword spotting)방식과 발화검증(utterance verification)방식으로 구분된다. 핵심어 검출 방식은 인식 대상 단어 외의 단어를 별도로 모델링하여 하나의 인식대상 단어처럼 사용하는 방식이고, 발화검증 방식은 각 음소마다 그와 유사한 anti-model을 작성한 후 정상적인 음소 모델과 anti-model과의 유사도를 비교하여 결정하는 방식이다. 본 연구에서는 독립단어 음성인식 시스템에 적용될 수 있는 발화 검증 방식에 의해 음성인식 거절 기능을 구현하였다. 특히 유사도를 결정함에 있어서 산술평균, 기하평균, 조화평균을 사용하고 각각을 비교하여, 기하평균을 사용하는 방식이 우수한 성능을 보임을 알 수 있었다. 음성의 신뢰도(confidence score)를 정규화하기 위해서 Sigmoid 함수를 사용하는데 이 함수의 가중치(weight) 상수의 변화에 대해 인식률을 비교함으로써 가장 적절한 가중치 상수값을 결정하였다. 그리고 유사음소집합(cohort set)에 대한 실험에서는 유사음소 집합의 크기가 클수록 더 좋은 성능을 보이는 결과를 얻었다. 음성인식 테스트 결과에서는 신뢰도 임계치 값을 구하고 이 값을 사용하여 인식률을 계산하였으며, 거절의 오류까지 포함된 음성인식률은 약 76%였다. 이 연구결과는 현재 한국통신에서 시험 서비스 중인 음성인식 증권정보 안내 시스템에 적용될 예정이다.

ABSTRACT

This paper presents the study on the rejection capability based on anti-phone modeling for vocabulary independent speech recognition system. The rejection system detects and rejects out-of-vocabulary words which were not included in candidate words which are defined while the speech recognizer is made. The rejection system can be classified into two categories by their implementation methods, keyword spotting method and utterance verification method. The keyword spotting method uses an extra filler model as a candidate word as well as keyword models. The utterance verification method uses the anti-models for each phoneme for the calculation of confidence score after it has constructed the anti-models for all phonemes. We implemented an utterance verification algorithm which can be used for vocabulary independent speech recognizer. We also compared three kinds of means for the calculation of confidence score, and found out that the geometric mean had shown the best result. For the normalization of confidence score, usually Sigmoid function is used. On using it, we compared the effect of the weight constant for Sigmoid function and determined the optimal value. And we compared the effects of the size of cohort set , the results showed that the larger set gave the better results. And finally we found out optimal confidence score threshold value. In case of using the threshold value, the overall recognition rate including rejection errors was about 76%. This results are going to be adapted for stock information system based on speech recognizer which is currently provided as an experimental service by Korea Telecom.

I. 서 론

최근들어 음성인식 기술이 발전함에 따라 음성인식을 이용한 다양한 음성대화 시스템들이 등장하고 있다. 이 시스템들은 사용자가 음성을 이용하여 사용할 수 있기 때문에, 사용자들로 하여금 자연스럽고, 편리한 인터페이스 방식을 제공한다는 장점을 지니고 있다. 그러나 이런 시스템들은 이미 시스템을 제작할 때 정해 놓은 인식대상 단어 이외의 단어(OOV : Out-Of-Vocabulary)들이 입력되었을 때 이를 처리할 수 없다는 단점을 지니고 있다. 즉 사용자는 미리 정해진 말만을 사용해야 하므로, 이 시스템을 사용하는데 있어서 상당한 제약을 받게 되는 것이다.

이런 문제점을 해소하기 위해 인식 거절(rejection)기능

* 한국통신 멀티미디어연구소 음성언어연구실
접수일자 : 1998년 7월 9일

이 연구되어 왔는데, 이는 인식대상 단어에 대해서만 인식을 하고, 그 외는 인식결과를 내지 않고 거절함으로써 시스템의 성능을 향상시키고자 하는 것이 목적이다. 인식 거절은 구현 방식에 따라서 핵심어 검출(keyword spotting) 방식[1]과 발화 검증(utterance verification) 방식[2]으로 구분된다.

핵심어 검출 방식이란 문법을 설계 할 경우 핵심어만 고려하고 그 이외의 단어는 garbage모델을사용하여 필요없는 단어를 제거하는 방식이다[3]. 제거하는 방법은 garbage모델의 likelihood값이 인식대상 핵심어의 likelihood 값보다 클 경우이다. 발화검증 방식이란 인식결과를 확인하는 과정이 추가 되며 이때 filler모델을 이용하는 방법이 사용되었다[4]. 그러나 이러한 filler모델은 단어를 기반으로 구성되었기 때문에 단어독립 음성인식을 위한 발화검증이 구현되기 위해서는 매 음소단위의 검증기능이 필요하다. 이를 위해서 반음소(anti-model)를 사용하는 방식이 제안되었다[5].

단어독립 음성인식 시스템이란 인식 대상 단어가 수시로 변화되는 경우에도 인식을 할 수 있는 시스템을 말한다. 다시 말해 인식 대상 단어가 새로 추가 되거나 변경되었다 하더라도 그 단어에 대해 새로이 훈련과정을 거치지 않고, 기존에 훈련된 정보를 바탕으로 인식하는 시스템이다. 이를 구현 할 경우 만약 음성인식의 단위가 단어라고 할 때, 실제로는 그보다 낮은 단위인 서브워드(sub- word), 음소(phoneme)나 그와 유사한 단위(PLU : Phoneme Like Unit)로 모델링을 하여 이 정보들에 의거하여 인식하는 방식을 사용한다. 즉, 훈련 시에 PLU 단위로 모델링을 하였다가 인식 시에도 단어 단위로 비교를 하지 않고, 각 인식대상 단어들의 구성 PLU 단위로 먼저 인식을 하여 이로부터 단어 단위의 인식 결과나 문장 단위의 인식 결과를 만들어 내는 것이다. 따라서 인식대상 단어가 변경되었다 하더라도, 이미 변경된 단어에 대한 PLU 단위와 정보는 이미 모델링된 상태이므로 이로부터 단어 단위의 인식 결과를 만들어 주는 과정만 변경해 주면 된다. 이 과정은 변경된 인식 대상 단어가 입력되면(여기에서 입력이란 훈련을 위한 음성 데이타의 입력이 아니라 변경된 인식대상 단어의 텍스트 입력을 말한다.) 이로부터 훈련과정 없이 단지 규칙에 의거하여 PLU 단위로 변경된다. 따라서 추가적인 음성 훈련이 없이도 단어독립 음성인식이 가능한 것이다. 단어독립 음성인식은 주로 인식 대상 단어들을 수시로 변경해야 할 경우에 유용하게 쓰인다. 예를 들어 인식대상 단어가 회사내의 부서명이라든가 혹은 증권시장에 상장된 회사명일 경우에 회사의 사정 혹은 주식 시장의 사정에 따라서 부서명과 회사명이 수시로 변경될 것이고, 이를 모두 수용하여 처리하려면 단어독립 음성인식 시스템이 요구된다.

본 논문에서는 반음소 모델을 이용한 발화검증 기능을 구현하였으며 특히 거절기능의 성능을 향상시키기위하여 신뢰도를 결정하기 위한 3가지 방식을 제안하였다. 또한 최적의 반음소 파라미터를 구하기 위하여 유사음소집합을 구하는 방식을 비교하였다. 먼저 2장에서는 단어독립 음성인식 시스템에 대해, 3장에서는 음성인식 거절기능의 두가지 구현방법에 대해 기술한다. 4장에서는 음성인식 거절 기능이 적용된 증권정보 안내 시스템에 대해 기술한다. 5장에서는 발화 검증 방식에서 신뢰도를 결정하기 위해 3가지 평균 산출 방법을 사용한 실험결과와 가중치 상수에 따른 신뢰도 변화에 대한 실험결과, 그리고 음성인식 테스트 결과를 나타낸다. 끝으로 6장에서는 결론을 맺고 향후 연구방향에 대해 논의한다.

II. 음성인식 거절 기능

음성인식 시스템은 인식대상 단위에 따라서 고립단어 인식 시스템과 연속음성 인식 시스템으로 구분된다. 전자는 사용자가 한 단어만을 말하거나 혹은 단어와 단어 사이에 명백한 구분을 둠으로써 시스템이 결국에는 단어 단위로 인식을 하는 시스템이다. 후자는 여러 단어나 문장을 자연스럽게 말을 하고, 시스템은 그 결과를 여러 단어나 문장 단위로 보여주는 것이다. 그러나 그 두가지 시스템 모두다 미리 정해 놓은 특정 인식 대상 단어만이 입력될 것이라는 가정 하에 음성인식 기능을 수행하며, 따라서 사용자가 실수로, 혹은 고의로 인식 대상 단어 외의 말을 해 버리면 인식대상 단어중의 하나로 인식결과를 보여주기 때문에 엉뚱한 말로 인식해 버리는 문제점을 지니게 된다. 다만 고립단어 인식 시스템의 경우는 그런 입력이 들어왔을때 그 단어에 대해서만 오인식을 하게 될 것이므로 그리 큰 문제가 되지 않는다고 볼 수도 있다. 그러나 연속음성 인식 시스템의 경우에는 음향학적 처리기(acoustic processor) 뿐만 아니라 언어학적 처리기(linguistic decoder)가 동작을 하게 되는데 이 언어학적 처리기에서 더 큰 문제를 야기시킬 수 있다. 즉, 사용자가 발화한 문장 중 오직 한 단어만 음성인식 대상 단어가 아니라 하더라도 그 단어가 오인식 됨으로 인해서 그 뒤에 발화된 단어에까지 영향을 미쳐 오인식률을 높이게 되므로, 더 치명적인 결과를 초래하게 된다.

따라서 어느 경우에든 인식 대상 단어 이외의 말에 대해 처리할 수 있는 기능이 요구되어 왔다. 즉 인식 대상 단어 외의 말이 입력되었을때 이를 다른 단어로 오인식하지 않고, 입력이 잘못되었음을 판단하는 음성인식 거절 기능에 대한 연구가 최근들어 활발히 진행되고 있다. 음성인식 거절 기능은 그 방식에 따라 핵심어 검출 방식과, 발화 검증 방식으로 구분된다.

2.1 핵심어 검출 방식(Keyword Spotting Method)

핵심어 검출 방식은 일반적인 음성인식에 사용하는 핵심어의 모델링 외에 비핵심어에 대해서도 모델링을 하였다가 이를 하나의 인식 대상 단어로 사용하는 방식이다. 따라서 대부분의 핵심어 검출 방식들은 핵심어 모델과 필러(filler) 모델을 사용하는 연결단어 인식 알고리즘을 기반으로 하고 있다. 여기서 필러 모델들은 핵심어에 해당되지 않는 음성구간들, 즉 비핵심어들과 비음성, 즉 묵음 또는 배경 잡음 구간들을 표현하는데 사용된다.

이 방식은 주로 핵심어 모델과 필러 모델 간의 구분이 비교적 명확한 경우에 좋은 성능을 발휘할 수 있다. 그래서 주로 핵심어는 인식대상 단어가 되고, 필러 모델은 그 이외의 말들로서 이 필러모델이 핵심어 모델과 유사하지 않으면서 그 외의 것들을 얼마나 잘 모델링하는가에 따라 성능이 좌우된다.

그러나 핵심어 검출 방식은 인식 대상 단어가 수시로 변경되는 단어독립 음성인식 시스템의 경우에 성능에 저하된다는 단점이 있다. 즉 핵심어와 비핵심어를 이미 확정한 상태에서 훈련을 마치고 이에 기반하여 음성인식 거절 기능을 수행하는데, 새로이 추가되거나 변경되는 단어가 이미 모델링된 비핵심어에 대한 유사도보다 핵심어에 대한 유사도가 높아야만 제대로 인식할 수 있다. 그러나 비 핵심어들이 핵심어 이외의 모든 말에 대해 모델링된 것이므로, 패턴 공간 상에서 좀더 일반적인 분포를 갖게 될 것이고, 따라서 새로 추가되거나 변경되는 단어들도 그 일반적인 패턴에 포함되어 비핵심어로 분류될 가능성이 높게 될 것이다.

핵심어 검출 방식은 미국 AT&T 사의 전화교환 업무 자동화 서비스에 1992년부터 적용되고 있고, 한국통신에서도 이미 핵심어 검출 방식을 이용하여 음성인식 거절 기능을 구현한 바 있다[6]. 이 시스템에서 핵심어는 인식 대상 단어이고, 비핵심어는 주로 끝점검출기(end point detector)의 오류로 인해 입력된 잡음이나 주변잡음들로 모델링하였다. 이 시스템의 인식 대상 단어는 상장된 주식의 회사이름으로서 주식 시장의 특성상 상장회사 이름이 변경되거나 추가, 삭제되는 경우가 빈번이 발생하고, 따라서 단어독립 음성인식 기술이 적용되고 있다. 기존의 시스템은 거절 방식을 구현하기 위해 핵심어 검출 방식을 사용하였으나, 본 연구에서는 단어독립 음성인식 시스템에서도 적용할 수 있는 발화 검증 방식의 거절기능을 구현하였다.

2.2 발화 검증 방식(Utterance Verification Method)

발화 검증 방식에서는 단어나 PLU 단위의 인식 결과를 받아들일 것인지(accept), 거절할 것인지(reject)를 결정하는 검증과정이 이용된다. 그 결정은 인식 결과와 같이 얻어진 신뢰도(confidence score 또는 confidence measure)에 의거하여 이뤄진다. 여기서 신뢰도란 음성인식 결과에 대해서 그 결과가 얼마나 믿을 만한 것인가를 나타내는 척도로서, HMM 모델의 Viterbi 탐색 결과 수치와는 다른 것이다. Viterbi 탐색 결과 수치는 어떤 단어나 음소에 대한 단순한 유사도를 나타낸다. 그러나 이 신뢰도란 인식된 결과인 음소나 단어에 대해서, 그 외의 다른 음소나 단어로부터 그 말이 발화되었을 확률에 대한 상대값을 말한다. 따라서 다른 말에 대한 그 말의 상대적 유사도라고 볼 수 있으며 이를 위해서는 각 음소나 단어에 대해서 가장 혼돈하기 쉬운 유사한 것들을 찾아서 그에 대한 HMM 모델을 만들어야 하며 이를 anti-model이라고 한다. 그리고 그 신뢰도 값이 정해진 어떤 임계치보다 클 경우에는 그 인식결과를 받아들여서 그 결과대로 인식했다고 보는 것이고, 반대로 작을 경우에는 그 결과를 신뢰할 수 없으므로 거절하게 되는 것이다.

발화 검증 문제를 통계적인 가설 테스트의 관점에서 수식화하여 나타내 보자. 우선 어떤 O를 실제 음성의 관측 세그멘트(segment)라 하면 음성인식 과정에서 O가 입력되었을 때는 크게 두가지의 가정이 가능하다. 즉, 그 O가 실제 어떤 음성 세그멘트 k로부터 발화되었을 것이라 가정하는 것이 가능한데 이를 null hypothesis라 하고 H_0으로 표현한다. 반면, 그 O가 실제 음성 세그멘트 k가 아닌 다른 유사한 음성에서 발화되었을 것이라 가정할 수 있는데 이를 alternative hypothesis라고 H_1로 표현한다. 그러면 주어진 테스트 세그멘트 O에 대해 발화 검증 과정은 null hypothesis에 대한 확률과 alternative hypothesis의 확률을 비교하여 null hypothesis에 대한 확률이 크면 이를 인식하고 아니면 거절하는 것이다.

$$P(O|H_0) > P(O|H_1) \tag{1}$$

위 식을 Bayes rule에 의해 다시 쓰면

$$P(H_0|O)P(H_0) > P(H_1|O)P(H_1) \tag{2}$$

$$\frac{P(H_0|O)}{P(H_1|O)} > \frac{P(H_1)}{P(H_0)} \tag{3}$$

이 된다. 여기서 $P(H_0|O)$는 HMM 모델 λ_k에서 O가 관측될 확률이고, $P(H_1|O)$는 그와는 다른 모델에서 O가 관측될 확률이다. 이 실험에서는 H_1을 모델링 하기 위해 각 음소마다 가장 유사한 음소들, 즉 cohort set을 구하여 이를 HMM 파라미터로 훈련하였으며 이렇게 훈련된 HMM 파라미터를 anti-model이라고 하고 $\lambda_{\bar{k}}$로 표현한다. 위 식에 log를 취하면 log-likelihood가 되는데 이를 $LLR_k(O, \lambda_k)$ 또는 줄여서 LLR로 표현한다[7].

$$LLR_k(O, \lambda_k) = \log P(O|\lambda_k) - \log P(O|\lambda_{\bar{k}}) \tag{4}$$

그리고 log-likelihood 값이 너무 큰 범위에서 나타나지 않도록 정규화시켰는데 이 정규화 함수로 Sigmoid 함수를 사용하였으며, 최종적인 신뢰도는 다음의 식에 의해 계산된다[8].

$$f(LLR) = \log \frac{1}{1+\exp(-\alpha \cdot LLR)} \tag{5}$$

발화 검증 방식은 검증과정을 인식과정과 동시에 수행하는가 아니면 인식과정이 끝난 후에 하는가에 따라 one-pass 또는 two-pass 구조로 분류한다[2]. Two-pass 구조는 기존의 인식과정 후에 그 인식결과를 검증하는 과정을 순차적으로 수행하기 때문에 인식결과에 대해 검증만 할 뿐 디코더(decoder)에 의해 생성된 결과를 수정할 수 없다

는 단점이 있다. 이에 반해 one-pass 구조는 신뢰도가 음성인식 디코딩의 과정 중에 사용된다. 그러므로 인식결과인 Viterbi score와 검증 결과인 신뢰도를 모두 고려한 최적의 탐색 경로를 찾아준다는 장점이 있다. Lleida와 Rose는 Viterbi 탐색 알고리즘을 수정하여 신뢰도를 고려한 수정된 Viterbi 탐색 알고리즘을 제안하였다[9].

III. 거절 기능을 갖는 기본 시스템

3.1 기본 시스템 구성

기본 시스템 구성을 위하여 one-pass 구조를 사용하기에 앞서 two-pass 구조를 사용하였다. One-pass 구조란 음성인식 기능과 검증기능이 동시에 검색이 되도록 하는 시스템이며[10], two-pass 구조란 인식가의 후처리 방식으로 검증기능을 구현하는 방법이다[2]. 그러므로 two-pass 구조는 기존에 구현되어 있던 시스템을 크게 수정하지 않고 추가로 검증과정만을 구현하여 사용하기 때문에 구현에 소요되는 시간을 단축시킬 수 있는 장점이 있다.

거절 기능을 갖는 음성인식 시스템의 구성도는 그림 1과 같다. 먼저 음성이 입력되면 끝점 검출기에 의해 음성구간만 검출된다. 검출된 음성은 특징 추출과정을 거치고, Viterbi 탐색 알고리즘에 의해 인식과정이 수행된다. 검증과정에서는 인식된 후보 단어들의 음소열에 대해 anti-model과의 LLR 값을 구해 그 단어의 신뢰도 값을 결정해 낸다. 그러면 그 신뢰도 값을 다시 임계치와 비교하여 신뢰도가 크면 그 인식단어로 인식하고 아니면 거절하게 된다.

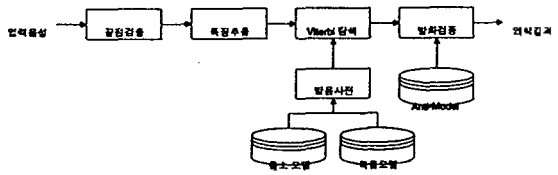

그림 1. 거절기능을 갖는 음성인식 시스템의 구성도
Fig. 1. The overview of the speech recognition system with rejection capability.

3.2 특징 추출

음성신호는 8KHz로 샘플링되고 전달함수가 $1-0.95z^{-1}$인 1차 디지털 필터로 pre-emphasis된다. 이 음성에 대해 매 10msec 단위로 LPC(Linear Predictive Coding) 분석이 행해지고, 주변잡음에 강하도록 가중치 함수에 의해 변환된다. 사용되는 특징은 다음과 같이 4종류, 총 38차의 벡타가 된다.

1. 12차의 LPC cepstrum
2. 12차의 LPC cepstrum 차이(delta cepstrum)
3. 12차의 LPC cepstrum 2차 차이(delta-delta cepstrum)
4. 2차의 Power 차이 및 power의 2차 차이(delta power and delta-delta power)

각 벡타는 훈련과정에서 구한 4종류의 VQ(vector quantization) 코우드 북을 사용하여 벡터 인덱스로 표현된다. 3개의 코우드 북은 256개의 코우드 워드로, 마지막 특징 벡터는 64개의 코우드 워드로 구성된다.

3.3 음소 HMM 모델

기본 시스템은 이산(discrete) 확률정보를 사용하는 HMM 인식 시스템이며 음소 단위로 HMM 파라미터를 추출한다. 본 논문에서는 62개의 문맥 독립음소(context independent phoneme)를 사용하고 이를 기준으로 unit reduction rule[11]에 의해 문맥 종속 음소(context dependent phoneme)를 생성한다. 그림 2에 음소 HMM 모델이 그려져 있다. 이 모델은 7개의 state와 12개의 transition으로 구성되며 관찰 확률은 매 transition마다 출력된다. 관찰 확률 분포는 B, M, E의 3가지 분포로 tying 시켰다.

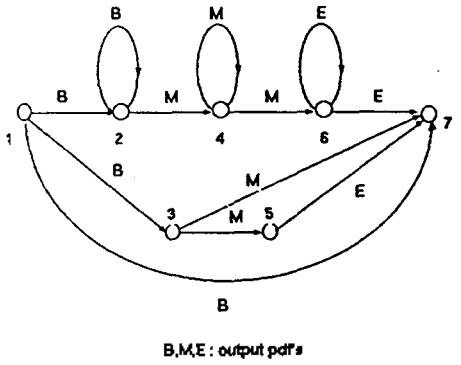

그림 2. 음소 HMM 모델
Fig. 2. The topology of a phoneme HMM model.

3.4 Anti-model을 위한 cohort set의 결정

정의된 문맥 독립 음소 각각에 대해서 가장 유사한 음소들, 즉 cohort set을 결정하기 위해서는 음소인식기가 구현되어야 한다. 이 음소 인식기는 음소 단위로 인식을 해야 하기 때문에 입력된 음성을 음소단위로 분할하는 과정이 필요하다. 이를 위해서 입력된 음성에 대해 HMM 파라미터를 이용하여 자동으로 해당음소 단위로 분할하는 자동음소 분할기를 구현하였다. 즉 단어 단위로 입력된 음성으로부터 훈련과정을 거쳐 저장된 HMM 파라미터에 의거하여 입력 음성을 그 구성 음소 단위로 분할하여 저장하고, 이를 4 : 1의 비율로 훈련용과 테스트용으로 구분하였다. 그래서 62개의 음소에 대해 음소 인식률을 구해서 각 음소들마다 어떤 음소로 오인식 되는 지를 구하였다.

IV. 실험결과

4.1 데이터베이스

음성인식 증권정보 안내 시스템은 상장된 회사명과 기타 단어들을 포함하여 총 1,062개의 단어를 인식할 수

있는 고립단어 인식 시스템이다. 이를 훈련시키기 위한 훈련 데이터는 총 62,717개이고, 테스트 데이터로는 9,751개를 사용하였다. 테스트 데이터에는 총 2,089개의 잡음 데이터(여기서 잡음 데이터란 인식대상 단어가 아닌 모든 발화를 의미한다.)가 포함되어 있다. 이 수치는 음성인식 증권정보 안내시스템의 시험서비스 중 수집된 데이터를 분석해 본 결과 수집된 총 발화 중 약 20%가 잡음 데이터였기 때문에 이와 유사하게 잡음의 비율을 맞추어 놓은 것이다.

4.2 평균산출 방법에 따른 변화

본 시스템의 경우 고립단어 인식 시스템이기 때문에 인식 결과가 단어단위로 나오는데, 실제로는 각 단어의 구성 음소 단위로 해당 음소 모델과 anti-model과의 확률값을 비교하게 된다. 그러나 인식 단위가 단어이기 때문에 신뢰도도 단어 단위로 산출해야 하고, 따라서 각 단어마다 그 구성 음소들의 anti-model과의 차이를 평균내어서 이를 단어 단위의 신뢰도로 사용한다.

평균을 내는 과정에서 우리는 3가지 평균 산출 방법을 사용하였다. 즉 산술 평균(arithmetic mean), 기하 평균(geometric mean), 조화 평균(harmonic mean)을 사용하여 각각의 경우에 신뢰도 값이 어떻게 나타나는가를 알아보았다. 그리고 이 신뢰도의 변화에 따른 성능을 측정하기 위해 신뢰도의 임계치(threshold)를 증가시켜 가며 오류가 어떻게 변화되는가를 그림 3에 나타내었다. 여기에서 오류는 잘못된 입력을 올바른 단어로 accept한 경우(false alarm)와, 올바른 입력에 대해 거절(reject)한 경우(false reject)를 모두 포함한 결과이다.

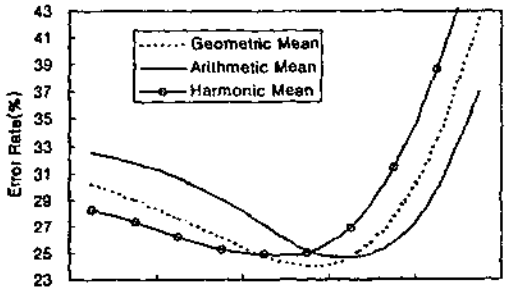

그림 3. 평균 산출 방법에 따른 신뢰도 임계치 대 오류율 그래프
Fig. 3. The error rate with regard to threshold when the three kinds of means are used.

그래프에서 알 수 있듯이 조화평균을 사용한 경우는 오류율이 가장 높게 나타났고, 산술 평균과 기하 평균을 사용한 경우는 최소 오류를 보이는 값은 유사하지만 기하 평균을 사용한 경우가 약간 오류율이 더 낮았다. 그리고 기하 평균을 사용한 경우에 그래프가 좀 더 완만한 곡선을 그렸고, 이는 임계치의 변화에 덜 민감하다는 것을 나타낸다. 다시 말해서 임계치가 약간 잘못 설정된다 하더라도 성능이 크게 나빠지지 않는다는 것이고 따라서 기하 평균을 사용한 경우가 가장 좋은 결과를 보임을 알 수 있었다.

4.3 가중치 상수에 따른 변화

두번째 실험으로는 *LLR* 값의 출력에 Sigmoid 함수를 가중치 함수로 사용한 경우에 가중치의 기울기 상수 α를 조절해 가며 적절한 값을 찾아보았다. 즉 가중치 상수 α의 값을 몇개 설정한 후 각각의 경우에 신뢰도 임계치 변화에 따른 오류율을 그래프로 표현하여 그림 4에 나타냈다. α값이 너무 작을 경우(0.1 또는 0.5의 경우)에는 임계치 변화에 너무 민감하게 반응하고, 반대로 너무 클 경우(3.0의 경우)에는 너무나 완만한 곡선을 보이게 되어 오류가 최소가 되는 임계치의 위치가 명확하게 나타나지 않는 결과를 보였다. 결국 α가 1.0일 경우에 적절한 임계치 범위를 뚜렷이 보이면서도 곡선이 어느 정도 완만하게 되어 가장 좋은 성능을 보였다. 따라서 가중치 상수 값은 1.0 정도의 값으로 설정하는 것이 바람직함을 알 수 있었다.

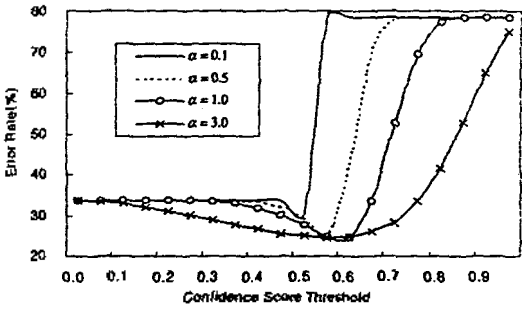

그림 4. 가중치 상수값의 변화에 따른 임계치 대 오류율 그래프
Fig. 4. The error rate with regard to confidence score threshold when various slope constants are used.

4.4 유사음소집합의 크기에 따른 변화

다음 실험으로 유사음소집합의 크기에 따른 오류율 변화에 대해 살펴보았다. 일반적으로 유사음소집합이 많을수록 반음소가 잘 모델링되지만, 유사음소집합의 크기가 너무 크게 되면 훈련 데이터량이 너무 많아지는 단점이 있다. 유사음소집합의 크기를 결정함에 있어서 두가지 방법을 사용하였다. 우선 첫째로 유사음소집합의 개수를 얼마로 고정시켜서 실험을 하였다. 그림 5에 보인 바와 같이 유사한 음소 개수를 상위 몇 개로 정하여 상위 10개, 30개, 60개로 테스트하였다. 이 실험결과 상위 60개로 한 것이 가장 좋은 결과를 보였다. 다음으로 상위 몇 %에 해당하는 음소들로 한정하여 테스트하였다. 그림 6에 보인 바와 같이 상위 50%, 상위 70%, 상위 90%에 해당하는 음소들로 테스트한 결과 상위 90%에 해당하는 경우가 가장 좋은 결과를 보였다. 이 두가지 실험에서 유사음소집합의 크기는 가능한 크게 하는 것이 좋다는 결론을 얻었다.

4.5 음성인식 테스트 결과

위의 실험에 의거하여 기하 평균을 사용하고 Sigmoid 함수의 가중치 상수를 1.0으로 사용하여 최적의 신뢰도 임계치 값을 결정하였다. 최적의 신뢰도 임계치 값은 0.575

였으며, 이 값을 사용하였을 경우의 실험결과는 표 1과 같다. 잘못 거절된 것까지 포함한 전체의 데이터에 대한 음성인식 결과는 75.96%로 나타났다.

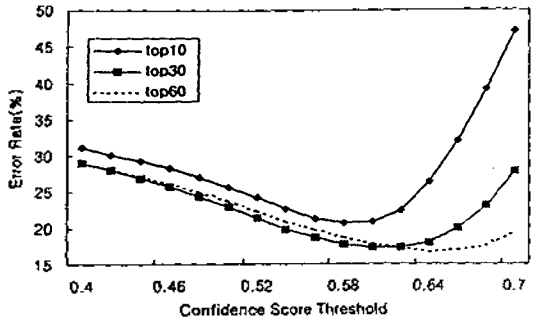

그림 5. 상위 N개 유사음소집합에 대한 테스트 결과
Fig. 5. The error rate with regard to confidence score threshold when top N phonemes are used.

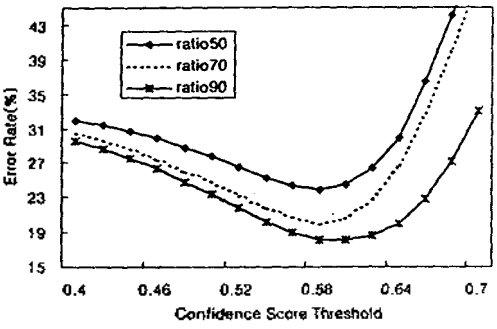

그림 6. 상위 N% 유사음소집합에 대한 테스트 결과
Fig. 6. the error rate with regard to confidence score threshold when top N% phonemes are used.

표 1. 음성인식 거절기능의 성능평가
Table 1. The rejection performance when the best parameters are used.

CA	CR	FAI	FAO	FR	2nd	TOTAL (NOISE)	REJECT RATIO(%)	REC RATE 1(2) (%)	TR_REC RATE(%)
6,171	1,236	879	847	618	408	9,751 (2,089)	6.34 %	81.10 (85.57)	75.96

이 표에 대한 설명은 다음과 같다.

CA : Correctly Accepted for Keyword, 즉 인식대상 단어를 제대로 accept한 경우
CR : Correctly Rejected for Noise, 즉 잡음에 대해 reject한 경우
FAI : False Accepted In Grammar Word(= Keyword), 즉 인식 대상 단어로 accept는 했지만 잘못 인식한 경우(Top 1으로 인식하지 못한 경우)
FAO : False Accepted Out of Grammar Word(= Noise), 즉 잡음인데 accept한 경우
FR : False Rejected for Keyword, 즉 인식대상 단어를 말했는데 reject한 경우
2nd : 두 번째 인식 후보로 맞게 인식된 경우
REJECT RATIO : 잘못 거절된 비율 = FR/TOTAL * 100
TOTAL : Noise를 포함한 총 테스트 데이터 개수 = CA + CR + FAI + FAO + FR
NOISE : 테스트 데이터 중 인식대상 단어가 아닌 모든 단어의 갯수
REC RATE 1 : Top 1에 대해 FR을 제외한 인식률 = 100 * (CA + CR) / (TOTAL − FR)
REC RATE 2 : Top 2에 대해 FR을 제외한 인식률 = 100 * (CA + CR + 2nd) / (TOTAL − FR)
TR_REC RATE : FR까지 포함된 실제 인식률 = 100 * (CA + CR) / TOTAL

V. 결 론

본 연구에서는 단어독립 음성인식 시스템을 위하여 반음소 모델링을 이용한 음성인식 거절기능에 대해 기술하였다. 음성인식 거절 기능은 음성 인식기를 제작할 때 정해놓은 인식대상 단어 이외의 단어가 입력되었을 때 그 단어가 인식할 수 없는 단어임을 알려주는 기능이다. 이 기능의 구현 방식은 핵심어 검출 방식과 발화검증 방식이 있다. 본 연구에서는 발화검증 방식을 사용하였으며 단어독립 음성인식 시스템에 적용이 가능 하도록 반음소 모델을 사용하는 방식을 구현하였다 특히 반음소 모델과의 유사도를 결정함에 있어서 산술평균, 기하평균, 조화평균의 사용방법을 제안하였으며, 성능을 비교한 결과 기하 평균을 사용하는 방식이 우수한 성능을 보임을 알 수 있었다. 또한 반음소모델 파라미터를 구하는데 필요한 유사음소집합을 구하는 방법을 비교하였다. 그 결과 유사음소집합의 크기는 가능한 크게 하는 것이 좋다는 결론을 얻었다.

이 연구결과는 현재 한국통신에 개발한 단어독립 음성인식 시스템인 증권정보 안내 시스템에 추가로 구현될 예정이다. 또 현재 음성인식과 검증 과정을 동시에 수행하는 one-pass 구조에 대해서도 구현이 진행되고 있다.

참 고 문 헌

1. R. C. Rose, "Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition," Computer Speech and Language, 9(9) : 309-333, 1995.
2. R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword in subword based speech recognition," IEEE Trans. on Speech and Audio Processing, Vol. 4, No. 6, pp. 420-429, Nov. 1996.
3. M. Rahim, et al., "Rpbust utterance verification for connected digits recognition", Proc. of ICASSP'95, pp. 285-288.
4. M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting", Proc. of ICASSP'95, pp. 297-300.

5. R. Sukkar, et al., "A vocabulary independent discriminatively trained method for rejection of non-keywords in subword based speech recognition," Proc. of EUROSPEECH'95 pp. 1629-1632.
6. 구 명완, "신경망을 이용한 음성인식 거절기능 구현," 제 13 회 음성통신 및 신호처리 워크샵, 제13권 1호, pp. 207-211, 1996.
7. Carmen Garcia-Mateo, C.-H. Lee, "A study on subword modeling for utterance verification in Mexican Spanish," Proc. on IEEE Workshop on Speech Recognition and Understanding, pp. 614-621, 1997.
8. M. W. Koo, C.-H. Lee, B. H. Juang, "A new hybrid decoding algorithm for speech recognition and utterance verification," Proc. on IEEE Workshop on Speech Recognition and Understanding, pp. 303-310, 1997.
9. E. Lleida, R. C. Rose, "Efficient decoding and training procedures for utterance verification in continuous speech recognition," Proc. IEEE-ICASSP, pp. 507-510, 1996.
10. E.Lleida, R. C. Rose, "Likelihood ratio decoding and confidence measures for continuous speech recognition," Proc. of ICSLP'96, pp. 478-481.
11. C.-H. Lee, et al., "Acoustic modeling of subword units for speech recognition," Proc. on IEEE-ICASSP, pp. 721-724, 1990.

▲김 우 성(Woosung Kim)
1990년 2월 : 한국과학기술원 전산학과 졸업(학사)
1992년 2월 : 포항공과대학교 대학원 전자계산학과 졸업(석사)
1992년 2월～1998년 7월 : 한국통신 멀티미디어연구소 음성언어연구실 전임연구원
※주관심분야 : 음성인식, 언어모델, 자연언어처리

▲구 명 완(Myoung-Wan Koo)
1982년 2월 : 연세대학교 전자공학과 졸업(학사)
1985년 2월 : 한국과학기술원 전기및전자공학과 졸업(석사)
1991년 8월 : 한국과학기술원 전기및전자공학과 졸업(박사)
1996년 12월～1997년 12월 : Lucent Technologies, Bell Labs.
Multimedia Communications Research Laboratory, Visiting Researcher
1985년 4월～현재 : 한국통신 멀티미디어연구소 음성언어연구실 실장
※주관심분야 : 음성인식, 합성, 자동통역 전화 연구 및 실용화