# 음소기반 인식 네트워크에서의 비인식 대상 문장 거부 기능의 비교 연구

김형태(강원대), 하진영(강원대)

<차 례>



<Abstract>

## Comparison Research of Non-Target Sentence Rejection on Phoneme-Based Recognition Networks

Hyung-Tai Kim, Jin-Young Ha

For speech recognition systems, rejection function as well as decoding function is necessary to improve the reliability. There have been many research efforts on out-of-vocabulary word rejection, however, little attention has been paid on non-target sentence rejection. Recently pronunciation approaches using speech recognition increase the need for non-target sentence rejection to provide more accurate and robust results.

In this paper, we proposed filler model method and word/phoneme detection ratio method to implement non-target sentence rejection system. We made performance evaluation of filler model along to word-level, phoneme-level, and sentence-level filler models respectively. We also perform the similar experiment using word-level and phoneme-level word/phoneme detection ratio method. For the performance evaluation, the minimized average of FAR and FRR is used for comparing the effectiveness of each method along with the number of words of given sentences. From the experimental results, we got to know that word-level method outperforms the other methods, and word-level filler mode shows slightly better results than that of word detection ratio method.

* Keywords: Sentence rejection, Filler model, Word/phoneme detection ratio, Speech recognition HMM

# 1. 서 론

최근 음성 인식 기술이 발전함에 따라 좀 더 자연스럽고, 편리한 인터페이스 방식의 음성 인식 시스템이 등장하고 있으나, 시스템 제작 시 정해 놓은 음성 문장 데이터외의 다른 데이터들이 입력되었을 때는 이를 적절히 처리하기 어려운 경우가 많이 있다[1]. 한 예로 영어 발음 교정 시스템의 경우 사용자로부터 학습 대상 문장이 아닌 음성이나 소음과 같은 기대하지 못한 소리가 입력되었을 때 이를 단순히 맞거나 틀린 문장이라고 판단하는 것 보다, 시스템이 거부함으로써 사용자가 제대로 된 문장을 재입력 할 수 있도록 하는 것이 더 편리하고 신뢰성 높은 시스템이라 할 수 있다. 또한 음성 인식에 의존하는 작업환경에서의 거부기능은 잘못된 패턴의 입력으로부터 시스템의 오작동을 방지 할 수 있다. 이러한 거부기능을 갖는 음성 인식 시스템의 성능은 신뢰도로 평가 할 수 있으며, 신뢰도가 높은 음성 인식 시스템은 단어 수준의 인식 시스템뿐만 아니라 문장수준을 인식하는 시스템에서는 필수적이라고 할 수 있다. 따라서 보다 높은 신뢰도를 갖는 음성 문장 인식 시스템을 개발하기 위하여 다양한 방법의 문장 거부 실험이 필요하다[2-9].

본 논문에서는 비인식 대상 문장 거부 기능을 구현하기 위하여 음소기반 인식 네트워크에서 필러 모델(filler model)과 단어/음소 검출률 모델(detection model)의 두 방법을 사용하여 문장 거부 성능 평가를 수행하였으며, 각각의 방법을 단어 수준과, 음소수준의 모델로 비교 실험 하였다.

필러 모델을 사용한 경우, 문장을 구성하고 있는 표준 음소로 된 단어들의 속성—유성자음(VC: Voiced Consonant), 무성자음(C: unvoiced Consonant), 모음(V: Vowel)—에 따라 각각의 필러 모델을 생성한 후 인식 대상 문장을 3가지 단위(단어 단위, 음소 단위, 문장 단위)의 필러 모델로 구성한다. 그리고 인식 네트워크에 병렬로 연결하여 입력 문장의 결과 속에 포함된 표준 모델과 필러 모델의 비율을 사용하여 문장에 대한 거부를 판단하였다. 단어/음소 검출률 모델의 방법은 직렬로 연결된 인식 네트워크를 통과한 인식 대상 문장 내 인식된 단어나 음소의 비율로 문장 거부를 판단하였다. 두 방법의 인식 성능을 평가하기 위하여 문장 길이에 따른 문장 거부의 성능을 FRR(False Rejection Rate: 제시된 문장이 입력되었음에도 이를 거부하는 오류)과 FAR(False Acceptance Rate: 제시된 문장 이외의 다른 문장이 입력되었을 때 거부하지 못하는 오류)의 평균을 최소화 하는 값을 각각 구함으로써 비교 · 분석하였다[10][11].

본 논문의 구성은 다음과 같다. 2장에서 음성 인식의 거부기능을 설명하고, 3장에서 제시한 방법을 사용한 문장 거부 네트워크의 구성을 보인 후, 4장에서 실험의 결과를 분석 후 5장에서 결론을 맺는다.

## 2. 음성 인식 거부 기능 연구에 대한 고찰

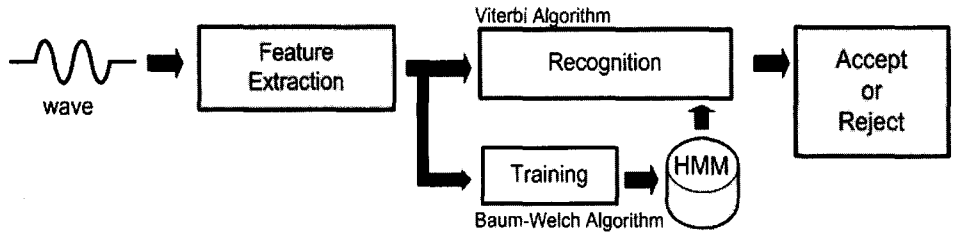
음성 인식 시스템은 발성 형태에 따라 고립 단어 인식 시스템과 연속 음성 인식 시스템으로 구분된다. 고립 단어 인식 시스템은 사용자가 한 단어만을 말하거나 단어와 단어 사이에 명백한 구분을 둠으로써 결국에는 단어 단위로 인식하는 시스템이다. 연속음성 인식 시스템은 여러 단어나 문장을 자연스럽게 말하고 시스템은 그 결과를 여러 단어나 문장 단위로 보여주는 것이다. 하지만 거부 기능이 없는 시스템은 모두 미리 정해진 특정 인식 대상의 단어와 문장들만이 입력될 것이라는 예상 하에 음성 인식을 수행하게 되므로, 사용자가 실수나 또는 고의적으로 인식 대상 단어외의 단어(out-of-vocabulary word)나 문장(non-target sentence)을 발음했을 경우에도 이를 인식 대상의 단어나 문장 중의 하나로 인식해 버린다. 따라서 신뢰성 있는 음성 인식 시스템의 구현을 위하여 인식 대상 문장 외의 단어나 문장이 입력되었을 때 이를 인식하려고 시도하는 대신, 입력이 잘못 되었음을 판단하여 이를 처리할 수 있어야 한다. 음성 인식 시스템의 성능은 오인식률 뿐만 아니라 거부율도 함께 고려해야 제대로 평가될 수 있다. 과거에는 단순한 음성의 인식결과에 따라 음성 인식 시스템의 성능을 평가하였으나, 음성 인식의 발달로 이러한 거부 기능은 음성 인식 시스템의 성능을 좌우할 수 있는 새로운 기준(신뢰도)이 되었고, 거부기능을 갖는 신뢰도 높은 음성 인식 시스템의 연구가 활발히 진행되고 있다. 한편 비인식 대상 단어 거부에 대한 연구와는 달리 비인식 대상 문장 거부에 대한 연구는 거의 이루어지지 않고 있다. 최근 주목 받는 지능형 로봇의 대화형 음성인식 인터페이스, 텔레매틱스, 홈 네트워크, 디지털 콘텐츠 검색 등의 차세대 음성인식 서비스의 질을 한 단계 높이기 위해서는 단어수준 음성인식과 함께 문장 수준의 음성인식이 필요하다.

문장 거부는 구현 방식에 따라 핵심어 검출(keyword spotting) 방식과 발화 검증(utterance verification) 방식으로 구분된다. 핵심어 검출 방식은 발화된 음성 중 인식 대상 핵심어만 추출하여 인식하고 나머지는 핵심어의 garbage 모델을 사용하여 제거한다. 발화 검증 방식은 문장의 인식결과를 받아들일 것인지(accept), 거부할 것인지(reject)를 결정하는 검증과정을 이용하며, 이에 대한 결정 검증과정은 신뢰도(confidence measure)에 의해 이루어지게 된다. 여기서 신뢰도란 인식된 결과인 음소 단어 문장에 대해서 그 외의 다른 음소 단어 문장이 발화 되었을 확률에 대한 상대 값이 된다. 즉 음성 인식 결과에 대해서 얼마나 믿을 수 있는가에 대한 척도를 나타내는 것이라 할 수 있다. 이런 음성 인식 엔진의 거절기능은 기본적으로 음소 및 신뢰도 점수(confidence score)에 근거하게 된다[6]. 본 논문에서는 인식 결과의 신뢰도를 높이기 위하여 특별한 핵심어의 유사도를 고려하지 않고, 발화된 음성의 목표 문장에 대한 인식된 표준음소의 비율과 필러 모델의 비율 이용한 문장 거부방법을 연구하였다.

# 3. 문장 거부 시스템

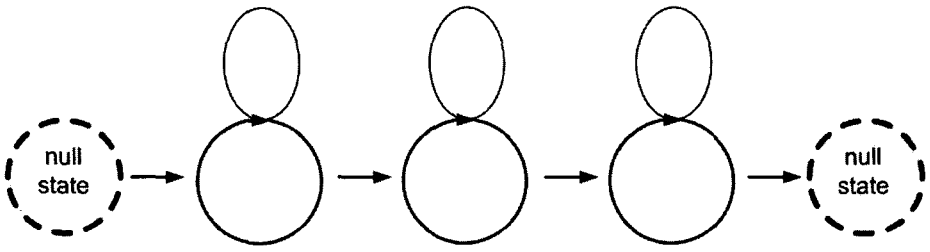
## 3.1. 시스템의 구성

본 논문에서는 비인식 대상 문장 거부를 위해, 최근 음성 인식 분야에서 우수한 성능을 보여 많이 사용하고 있는 HMM(Hidden Markov Model)을 사용하였다. <그림 1>은 전반적인 문장 거부 시스템의 구성을 보여준다. 보통의 음성 인식 시스템과는 달리 시스템의 출력은 2가지로 국한된다. 주어진 문장 발화가 입력되었다고 판단하느냐(Accept) 아니면 거부 하느냐(Reject)이다. 이러한 거부 시스템은 일반적인 음성 인식 시스템과는 달리 특정 문장 발화에 대해 발음의 정확도, 유창성, 억양 등을 평가하는 발음 교정 시스템에 사용될 수 있다[1].




<그림 1> 비인식 대상 문장 거부 기능 수행 흐름도

본 논문에서는 "Don't miss the bus."라는 문장을 사용하여 문장 거부 방법을 설명하기로 한다. 이 문장을 IPA(International Phonetic Alphabet: 국제 음성 기호)를 사용하여 발음을 표기하면 /d/ /óu/ /n/ /t/, /m/ /í/ /s/, /ð/ /ə/, /b/ /ʌ́/ /s/와 같이 나타낼 수 있다. <그림 2>는 음소 모델의 구조를 이용한 비인식 대상 문장 거부 네트워크 구성이다.




<그림 2> 음소 모델의 구조

## 3.2 필러 모델과 단어/음소 검출률

문장 거부를 수행하는 음성 인식 시스템에서의 신뢰도는 인식 결과를 받아들일 것인지 거부할 것인지 결정하게 된다. 여기서 신뢰도는 인식 대상 단어나 문장에 대한 HMM모델의 Viterbi 알고리즘의 결과에 대한 신뢰도가 아니라 인식 대상 문장에 대해 그 문장이 올바르게 발화되었을 확률에 대한 신뢰도를 의미한다. 이때 문장의 인식률과 거부율은 네트워크의 구성의 방법에 따라 달라질 수 있다. 본 논문에서는 필러 모델을 사용한 방법과 단어/음소 검출률을 이용하는 방법에 따라 각각의 인식 네트워크를 구성하였다.

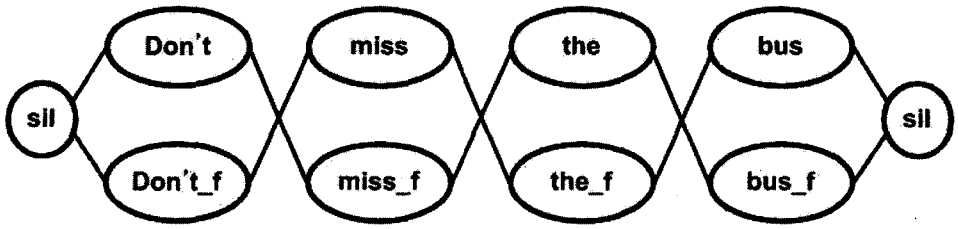
### 3.2.1. 필러 모델을 사용한 방법

필러 모델이란 인식 대상 단어 이외의 단어나 혹은 자동차 소리, 울음 소리등과 같은 음성을 모델링하기 위해 사용된다. 인식대상 단어 이외의 단어를 표현하기 위해서는 대표적인 감탄사나 많이 사용되는 단어를 독립 단어로 모델링하고, 자동차 소리, 울음소리, 숨소리 등과 같은 비 음성 데이터에 대해서도 각각 독립 단어로 모델링을 한다. 필러 모델링 기법은 고립단어 인식기나 핵심어 검출 시스템에서 인식 어휘의 거부 기능을 구현하기 위해서 별도의 필러 모델을 구성해야 한다[13]. 따라서 사용자로부터 발화된 음성이 인식기를 거쳐 나온 최종 결과는 인식된 단어의 표준 음소나 필러모델이 되며 이는 인식 대상 문장의 단어의 수와 일치하게 된다. 본 논문에서는 3가지 수준(단어수준, 단어별 음소수준, 문장전체)의 필러 모델을 만들어 이를 해당 네트워크에 연결하여 인식 실험을 하였다.

### 3.2.2. 단어 수준 필러 모델

문장 인식 네트워크에서 필러 모델은 인식되는 문장 내 단어 외의 모든 단어나 소음 등에 대하여 다양한 구성이 가능하며, 본 논문에서는 각각의 방법에 따른 3가지 방법의 필러 모델 인식 네트워크를 구성하였다.

문장 내 단어수준으로 구성한 필러 모델을 사용한 방법(방법 1)은 <그림 3>과 같은 음소기반 인식 네트워크로 구성할 수 있다. 대상 문장의 각 단어별마다 필러 모델을 만들어 인식 대상 단어와, 필러모델을 인식네트워크에 병렬로 연결하였다. 따라서 본 네트워크 인식 경로는 표준 음소로 구성된 인식 대상 단어와 필러모델로 가는 양자 택일형이며, 화자의 발음에 따라 달라질 수 있다. 단어 단위의 필러 모델을 사용하는 방법은 인식 대상 단어마다 해당하는 단어 단위의 필러모델이 존재하게 되므로 anti-keyword를 사용하는 방법과는 차이가 있다. <그림 3>의 Don't_f, miss_f, the_f, bus_f는 각각 Don't, miss, the, bus의 필러 모델이다.




<그림 3> 문장의 단어별 필러 모델 인식 네트워크 (방법 1)

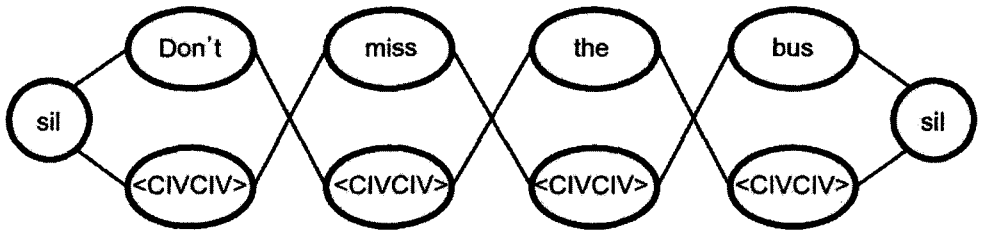
### 3.2.3. 단어별 음소 수준 필러 모델

문장 내 단어들을 구성하는 음소들을 이용한 방법(방법 2)은 <그림 4>와 같이 문장의 각 단어별 표준음소와 유성자음(VC), 무성자음(C), 모음(V)을 반복 인식할 수 있도록 병렬로 네트워크를 구성하였다. 음소 수준의 필러 모델은 문장의 각 단어의 음소에서 음성 신호의 특징 차이가 가장 큰 반대가 되는 발음을 사용하였다. <표 1>과 같이 유성자음(VC)은 무성자음(C)으로, 무성자음(C)은 유성자음(VC)으로 변환하였다. 또한 모음(V)의 경우 유성음으로 유성자음보다 무성자음이 더 반대의 성질을 갖게 되므로 모음은 무성자음으로 변환하였다.

<표 1> 문장의 음소 단위 필러 모델 변환

| |
|---|
| 무성자음(Unvoiced Consonant) → VC |
| 유성자음(Voiced Consonant) → C |
| 모음(Vowel) → C |

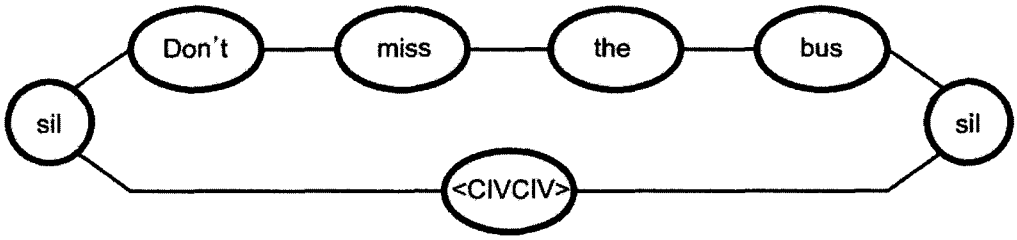
| 모델 타입 | 단어별 | 음소 단위 변환 |
|---|---|---|
| 표 준 | don't | /d/ /óu/ /n/ /t/ |
| 필 러 | don't_f | C C VC VC |
| 표 준 | miss | /m/ /í/ /s/ |
| 필 러 | miss_f | C C VC |
| 표 준 | the | /ð/ /ə/ |
| 필 러 | the_f | C C |
| 표 준 | bus | /b/ /ʌ/ /s/ |
| 필 러 | bus_f | C C VC |




<그림 4> 문장의 단어별 C|VC|V 반복 인식 네트워크 (방법 2)

### 3.2.4. 문장 수준 필러 모델

문장 전체에 대한 음소수준 필러모델을 이용한 방법(방법 3)은 인식 대상 문장 전체에 대하여 <표 1>의 유성자음(VC), 무성자음(C), 모음(V)을 반복하는 필러 모델을 사용하여 <그림 5>와 같이 구성하였다. 따라서 입력된 음성 문장에 대하여 인식된 문장 전체가 올바른 문장인지, 올바르지 않은 문장인지를 판단하도록 하는 문장 거부 네트워크를 구성하였다.




<그림 5> 문장 전체에 대한 C|VC|V 반복 인식 네트워크 (방법 3)

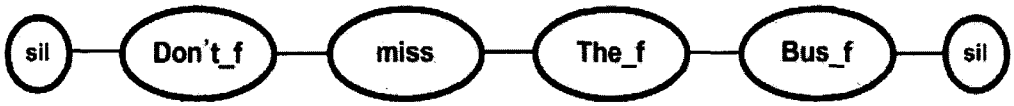
### 3.2.5. 필러 모델을 사용한 검증

필러 모델을 사용한 각각의 방법은 표준 발음 인식 네트워크에 필러 모델 단위를 연결하여 비인식 대상 문장에 대한 거부기능을 수행할 수 있다. 인식률과 거부율의 판단은 각각의 인식 네트워크를 거친 음성데이터에 대하여 표준 음소 모델과 필러 모델의 비율에 따라 결정된다.

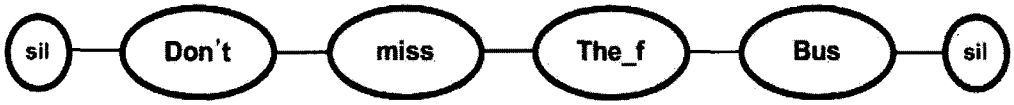
입력된 음성 문장이 각각의 네트워크에서 Viterbi 알고리즘을 이용한 최적 경로를 통과한 결과 데이터에 대하여 우리는 식 (1)에서와 같이 필러 모델 비율 ($ratio_f$)을 구할 수 있다.

$$ratio_f = \frac{number\ of\ filler\ models}{number\ of recognized\ models} \tag{1}$$

만약 입력된 음성 문장에 대해 대상 문장을 발음 하였다면 필러 모델 비율이 감소하게 되고, 대상 문장을 제대로 발음하지 않았거나, 전혀 다른 문장을 발음한다면 필러 모델 비율은 증가할 것이다. 따라서 문장 거부 시스템의 거부 기능의 수행은 문턱값($\theta_f$)에 의존하게 된다. 입력 데이터에 대하여 필러 모델의 비율이 문턱값보다 크다면($ratio_f > \theta_f$) 거부를 하고, 그렇지 않으면 제대로 인식되었다고 판단을 할 수 있다. <그림 6>은 필러 모델을 사용한 문장거부의 예를 보여주고 있다. <그림 6>의 (a)에서 총 4개 인식 대상 단어 중 인식 결과로 Don't_f, The_f, Bus_f 의 필러 모델이 존재한다. 따라서 문턱값($\theta_f$)보다 필러 모델의 비율이 크므로 거부하게 되고, 반대로 (b)에서는 올바른 인식으로 간주한다.




(a) Reject ($ratio_f > \theta_f$)




(b) Accept ($ratio_f \leq \theta_f$ )

<그림 6> 필러 모델을 사용한 문장 거부 예

또한 식 (2)를 이용하여 각 문장의 단어수별 FAR과 FRR의 평균을 최소화 할 수 있는 최적의 문턱값($\hat{\theta_f}$)을 구할 수 있다.

$$\hat{\theta_f} = \arg_{\theta_f}\min\left[\frac{FAR_{\theta_f} + FRR_{\theta_f}}{2}\right] \tag{2}$$

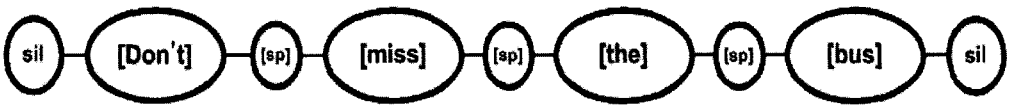
### 3.3. 단어/음소 검출률을 이용한 방법

비인식 대상 문장 거부를 위한 또 다른 방법으로, 단어/음소 검출률을 이용한다. 단어/음소 검출률 방법이란 입력된 음성 데이터에서 인식된 결과의 문장 내 포함된 단어나 음소의 검출된 비율에 의존하는 방법이다. 검출률을 이용한 방법은 앞서 제시한 필러 모델을 사용한 방법과 달리 별도의 필러 모델 훈련 생성이 필요 없고, 네트워크를 거친 입력 데이터에 대한 인식된 단어나 표준음소의 비율로 거부기능을 수행한다. 본 논문에서는 검출률 방법을 문장 내 단어수준의 검출률

방법과 단어의 음소수준의 검출률 방법으로 분류하여 각각에 대하여 실험 하였다.

### 3.3.1. 단어 수준 검출률

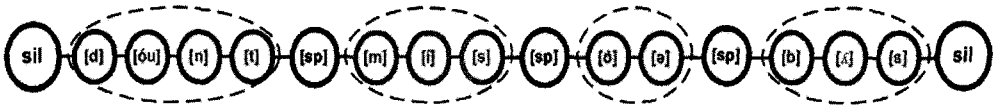
단어 수준 검출률을 이용한 방법은 <그림 7>과 같이 문장 내 인식 대상 단어 모델들을 [ ] -optional item-으로 처리하여 네트워크에 직렬로 연결하였다. 그리고 각 단어모델 사이는 sp(short pause)로 연결하였다. 이는 인식에는 크게 영향을 미치지 않으나 sp가 여러 개 연결되어 하나의 인식된 단어를 대신 할 수도 있다. 사용자로부터 발화된 입력된 음성은 네트워크를 거치면서 선택적으로 인식되어 분리된다. 따라서 앞서 제시한 필러 모델을 사용하는 방법과는 다르게 단어 수준 검출률을 이용한 방법은 음성 인식기에서 최종 인식된 단어들의 수가 인식 대상 문장의 단어들의 수와 다를 수도 있게 된다. 문장의 인식과 거부는 인식된 결과의 단어나 음소의 수에 의존하게 된다.




<그림 7> 단어 수준의 검출률 모델을 이용한 방법 (방법 4)

### 3.3.2. 음소 수준 검출률

음소 수준 검출률을 이용한 방법은 앞서 제시한 단어 수준 검출률을 이용한 방법과 유사하며, <그림 8>에서 보는 것과 같이 인식 대상 문장 속의 단어들을 단어 단위가 아닌 음소단위를 [ ] -optional item-으로 처리하여 네트워크에 직렬로 연결하였다. 따라서 단어 수준의 검출률을 이용한 방법과는 다르게 최종 인식된 단어들 속에서도 인식된 음소와 인식되지 않은 음소로 분류된다. 문장의 인식과 거부는 인식된 음소들의 비율로 문장의 인식과 거부를 판단할 수 있다.




<그림 8> 음소 수준의 검출률 모델을 이용한 방법 (방법 5)

### 3.3.3. 단어/음소 검출률을 이용한 검증

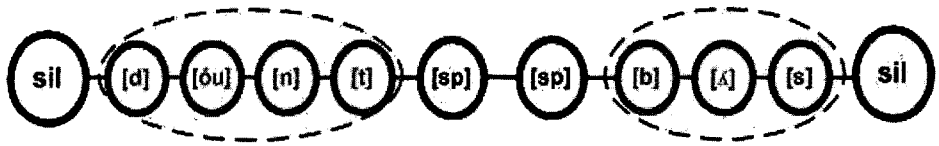
단어/음소 검출률을 이용한 방법은 인식 네트워크를 거친 문장의 결과를 이용한다. 인식 결과로 나온 문장은 인식된 단어/음소와 인식 되지 않은 단어/음소로 구분된다. 문장의 거부 기능은 결과로 나온 인식된 단어의 비율($ratio_{dw}$) 또는 음소의 비율($ratio_{dp}$)에 따라 판단된다. 문장 내 인식된 단어나 음소의 비율은 식 (3)을 사용하여 구할 수 있다.

$$ratio_{dw} = \frac{number\ of\ recognized\ models}{number\ of\ total\ words\ in\ a\ sentence} \tag{3}$$

$$ratio_{dp} = \frac{number\ of\ recognized\ models}{number\ of\ total\ phonemes\ in\ a\ sentence}$$

입력된 대상문장에서 결과로 인식된 대상 단어나 음소의 비율이 문턱값($\theta_d$)보다 높으면($ratio_{dw} \geq \theta_d$ 또는 $ratio_{dp} \geq \theta_d$) 이것은 올바른 문장으로 판단할 수 있고, 그렇지 않으면 거부를 할 수 있다. <그림 9>의 예를 보면 각각의 문장들은 주어진 인식네트워크를 통과하여 (a)와 (b)의 결과 로 나타나며, 이때 인식과 거부는 인식 대상 문장속의 올바르게 인식된 단어나 음소의 수에 따라 결정된다.


(a) 단어 검출률 방법의 예: Accept ($ratio_{dw} \geq \theta_d$)




(b) 음소 검출률 방법의 예: Reject ($ratio_{dp} < \theta_d$)

<그림 9> 단어/음소 검출률 모델을 이용한 문장 거부 예

단어/음소 검출률을 이용한 방법의 FAR과 FRR을 최소화 할 수 있는 검출률 모델의 최적의 문턱값($\hat{\theta}_d$)은 식 (4)에 의하여 구할 수 있다.

$$\hat{\theta}_d = \arg_{\theta_d}\min\left[\frac{FAR_{\theta_d} + FRR_{\theta_d}}{2}\right] \tag{4}$$

# 4. 실험 및 결과 분석

## 4.1. 실험 환경

비인식 대상 문장 거부 실험을 위하여 실험에 사용될 각 문장(4,120개)을 문장을 구성하는 단어의 수(2~6)별로 분류하였다. 그리고 분류된 문장에 필러 모델 방법을 사용하기 위하여 문장의 음소 단위별 필러 모델을 제작하여 필러 모델 인식 네트워크를 구성하였다. 그리고 검출률 방법을 실험하기 위하여 문장내의 모든 단어를 선택적으로 인식할 수 있도록 네트워크를 구성하였다. 음향 모델 훈련과 인식 실험은 HTK(Hidden Markov Model Toolkit) V.3.2.1을 사용하여 수행하였다.

본 실험에서 사용한 영어 음성 데이터베이스는 언어교육을 위한 영어 발음 교정용 음향 모델 생성을 목적으로, PC 환경에서 영어를 모국어로 사용하는 성인 400명이 문장을 발음한 영어 음성 DB를 사용하였다. 음성 데이터는 16KHz, 16bit, Mono, linear PCM으로 녹음되었으며, 남자 200명, 여자 200명이 각각 발음한, 총 4120개의 영어 문장을 사용하였다.

실험에 사용된 사전은 4.58MB의 크기를 갖는 표준 발음사전이며, CMU 사전을 근간으로 하여 만들었다. 음소 모델은 <표 2>와 같이 총 113개의 음소를 사용하였다. 또한 가우시안 믹스쳐(Gaussian Mixture) 7개의 Continuous density HMM을 사용하였다.

<표 2> 음소 모델

| 구 분 | | 음소 개수 |
|---|---|---|
| 표준 자음 | | 25개 |
| 표준 자음에 대한 변이음 | | 25개 |
| 모음 | | 14개 |
| 강세가 있는 모음 | 유성음으로 끝나는 경우 | 14개 |
| | 무성음으로 끝나는 경우 | 14개 |
| 단순모음 + 자음 | | 6개 |
| 한국어 자음/모음 | | 8개 |
| sp, sil, VC, C, V를 포함하는 특정 발성음 | | 7개 |
| 계 | | 113개 |

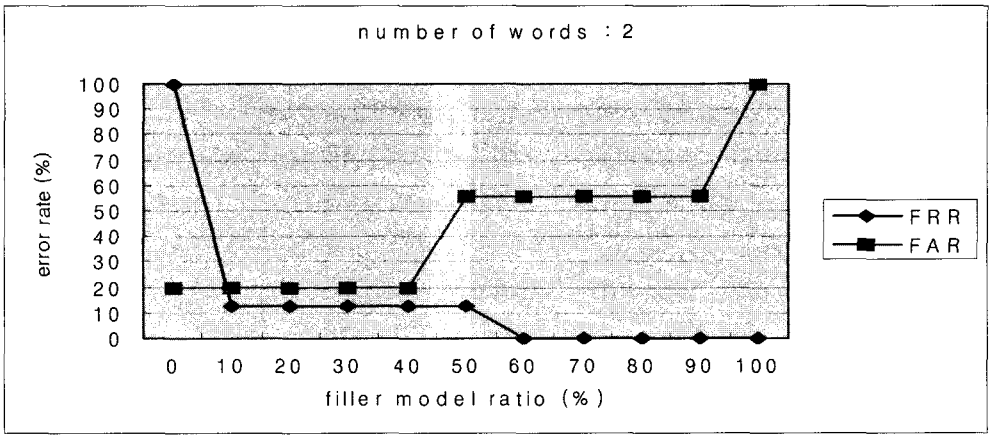
## 4.2. 실험 결과

비인식 대상 문장 거부 기능을 실현하기 위한 기본적인 방법은 음성 인식 네트워크에서 Viterbi 알고리즘으로 최적 경로를 구한 결과에 포함되어 있는 표준 모델 비율이다. 따라서 입력 음성에 대하여 네트워크에서 주어진 문장의 모든 단어가 모두 올바르게 인식이 되었다면 입력 음성은 주어진 문장을 발화한 것이라고 판단할 수 있고, 반대로 모두 인식 되지 않았다면 주어진 문장대신 다른 문장을 발화 하였거나 소음으로 판단하여 거부해야 할 것이다. 그러나 이것은 이상적인 결과이며 주어진 문장을 입력 하였음에도 이를 거부할 수도 있고, 다른 문장을 입력하였는데도 정상 인식된 결과가 나올 수 있다. 따라서 적절한 문턱값(threshold)의 설정이 필요하다. 문장 거부 시스템의 문턱값을 높이면 FAR이 커지고, 문턱값을 낮추면 FRR이 커지기 때문에 FAR과 FRR의 평균치를 최소화 할 수 있는 적절한 문턱값을 선택이 필요하다.

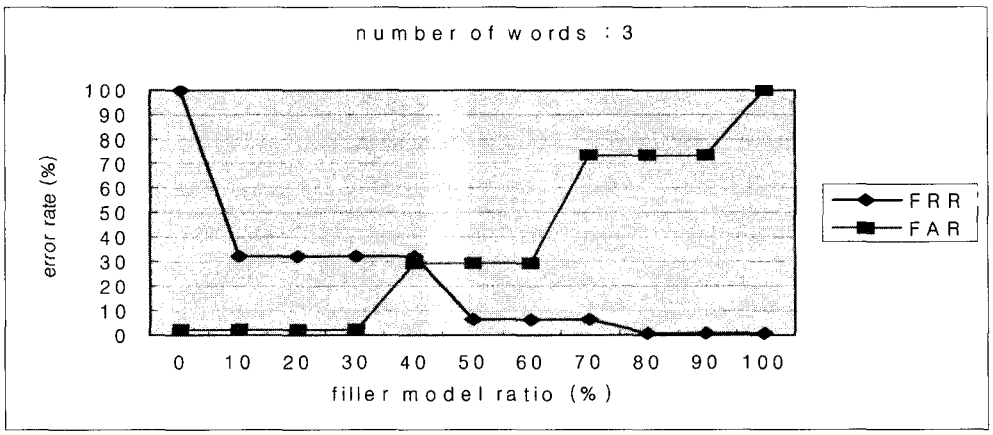
본 논문에서는 필러 모델을 사용한 3가지 방법(방법 1, 방법 2, 방법 3)과 단어/음소 검출률을 이용한 방법(방법 4, 방법 5)의 총 5가지의 실험을 하였다. 실험은 인식대상과 같은 음성 문장 데이터를 입력하였을 경우(FRR)와, 전혀 다른 음성 문장을 입력 하였을 경우(FAR)에 따른 문장 거부 오류율을 구하였으며 그 결과를 각각 Result01, Result02, Result03, Result04, Result05로 표시하였다. 또한 목표 문장의 문장 내 단어 수(2~6단어)에 따라 가변적인 길이에 따라 결과를 구분하였다.

### 4.2.1. 필러 모델을 사용한 방법

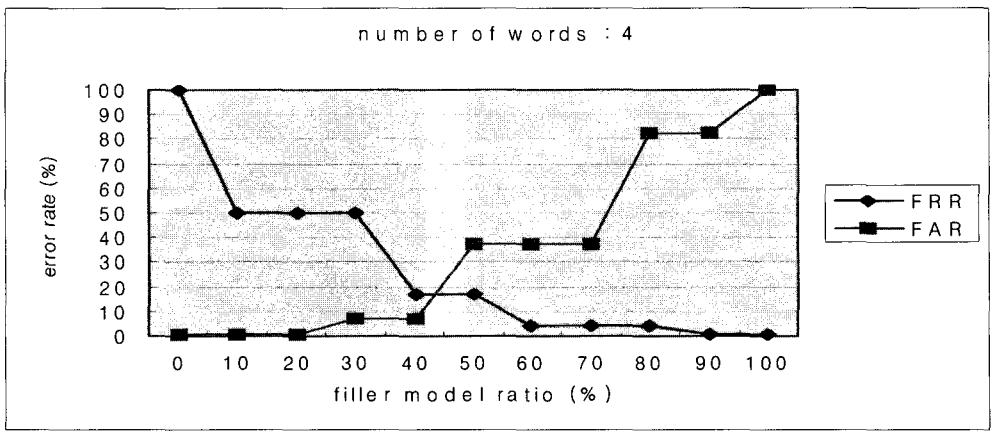
필러 모델을 사용한 방법으로 문장 거부 기능을 구현하는 방법은 앞서 제시한 3가지 방법을 이용하였으며, 각각의 방법에서 필러 모델 비율에 따른 오류를 구하고 이를 FAR과 FRR의 그래프로 나타내었다. <그림 10>의 그래프는 본 논문에서 제시한 문장의 단어별 필러 모델 인식 네트워크 실험(방법 1)의 결과인 백분율 누적값을 사용하여 인식 대상 문장의 단어 수에 따라 FAR과 FRR의 그래프로 나타낸 것이다.



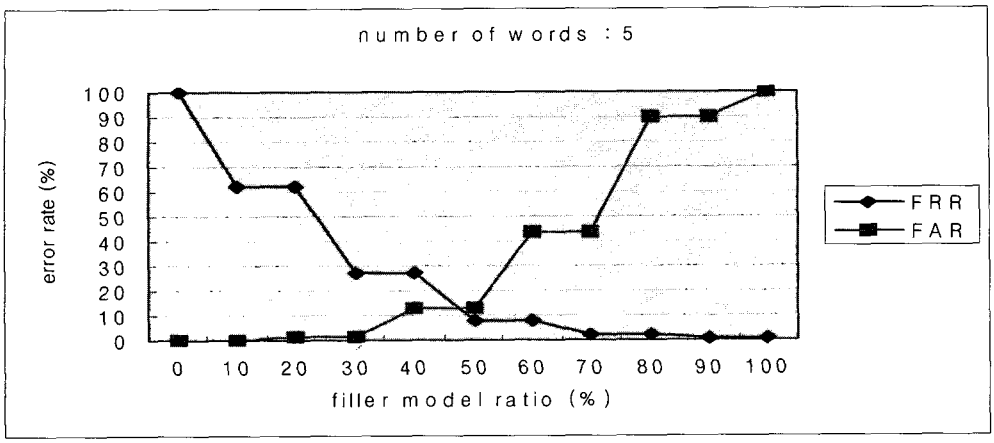

(a) 2단어



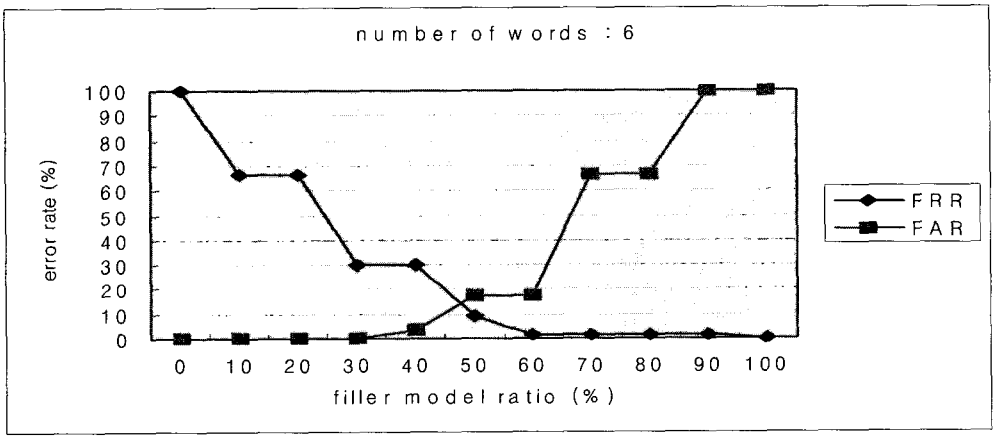

(b) 3단어




(c) 4단어



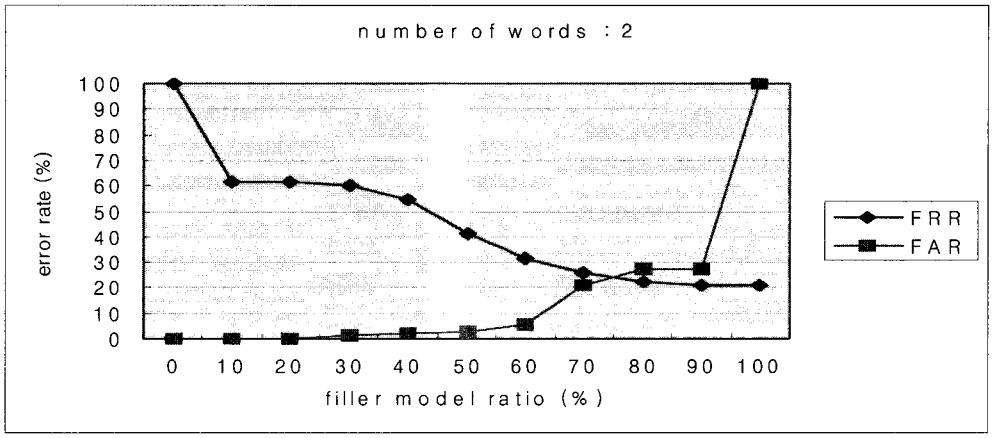

(d) 5단어



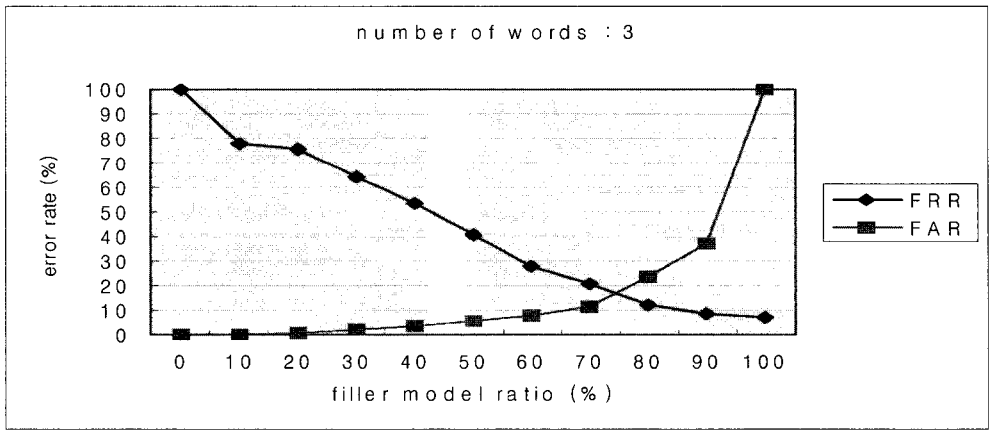

(e) 6단어

<그림 10> 문장의 단어수별 필러 모델 인식 네트워크에 따른 결과 (Result01)

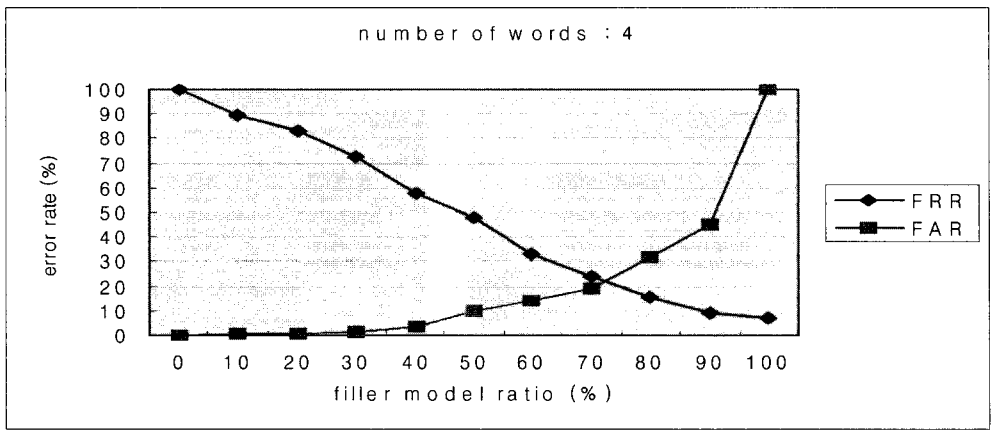
문장의 단어별 C|VC|V 반복 인식 네트워크 실험(방법 2)은 인식 네트워크의 결과의 백분율 누적값을 사용하여 문장의 단어수에 따라 분류하여 <그림 11>과 같이 그래프로 나타내었다.



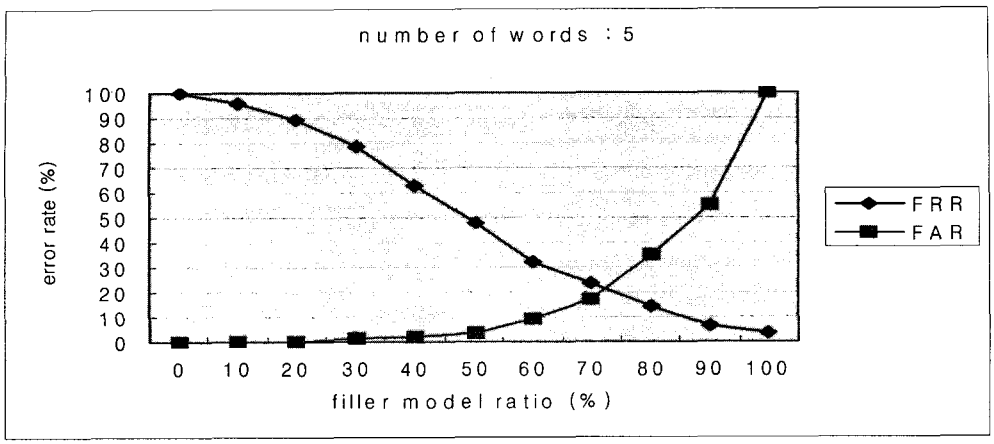

(a) 2단어



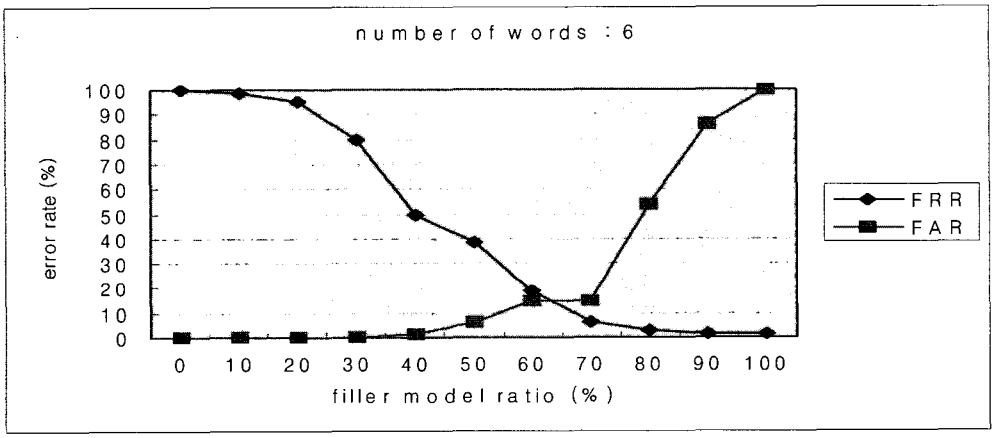

(b) 3단어




(c) 4단어




(d) 5단어



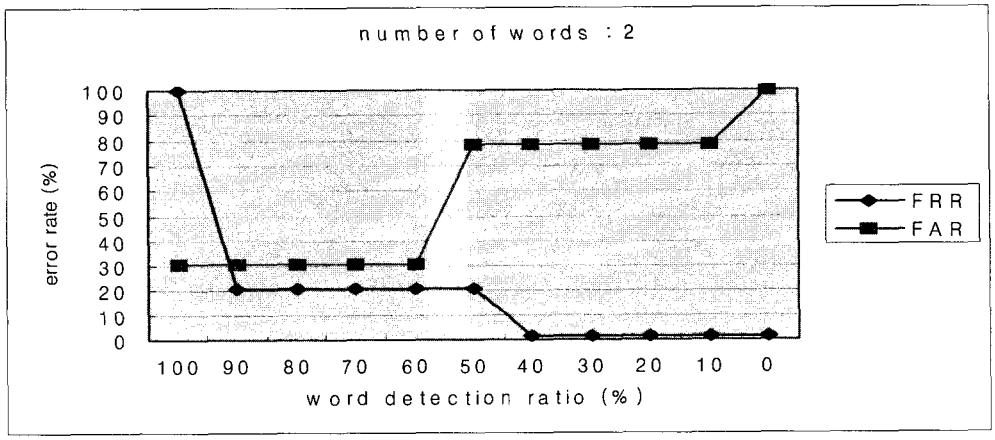

(e) 6단어

<그림 11> 문장의 단어수별 C|VC|V 반복 인식 네트워크에 따른 결과 (Result02)

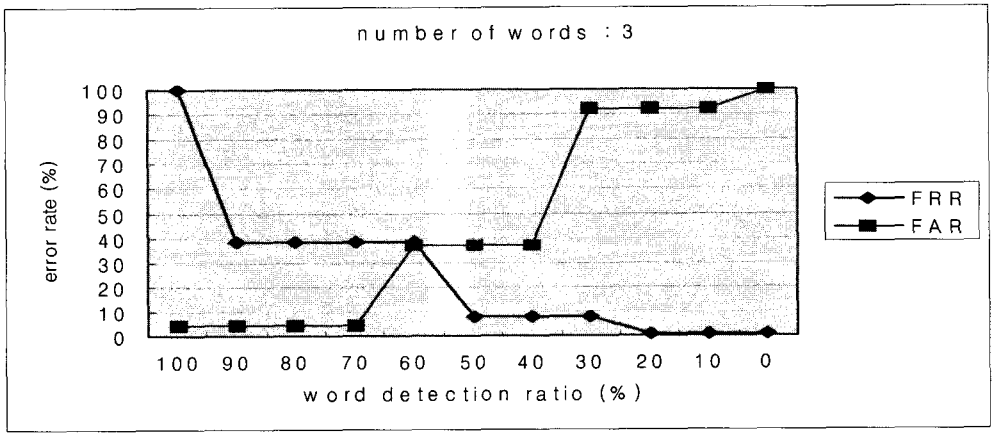
각각의 방법을 통한 실험 결과 단어수를 고려하여 가변 문턱값을 적용하였을 때 (방법 1)에서는 13.29%, (방법 2)에서는 19.23%, (방법 3)에서는 20.56%에서 오류의 비율이 최소가 되는 평균 FAR과 FRR을 구할 수 있었다. 실험 결과 (방법 1)의 단어별 필러 모델을 적용한 방법이 가장 좋은 결과를 얻을 수 있었으며, (방법 3)의 경우 문장의 인식 결과에 필러 모델이 한 개라도 있으면 에러를 보이기 때문에 가장 좋지 않은 인식 결과를 얻었다.

### 4.2.2. 단어/음소 검출률을 이용한 방법

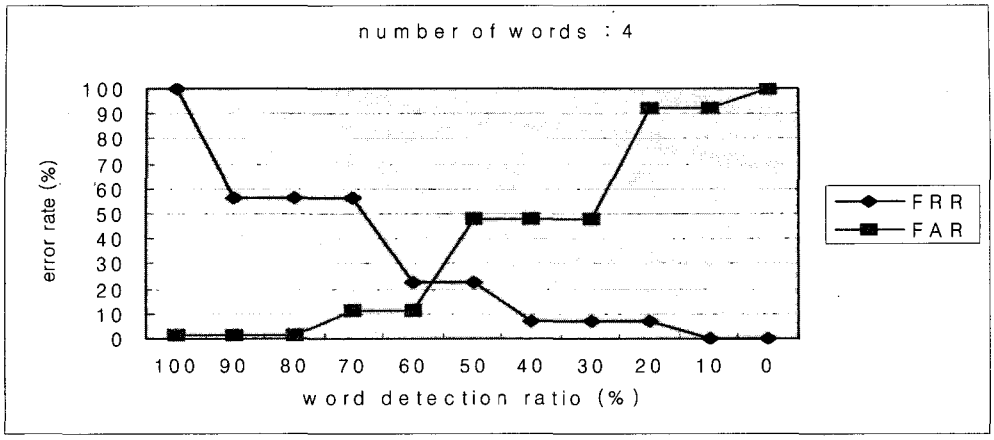
단어/음소 검출률을 이용한 문장 거부 실험은 (방법 4)와 같이 별도의 필러 모델을 생성하지 않고, 문장 내 포함된 인식된 단어만을 이용하여 단어/음소 단위의 실험 하였다. <그림 12>는 단어 수준 검출률을 이용한 방법(방법 4)의 네트워크에 인식 대상 문장을 입력하였을 때 문장 내 포함된 인식된 단어수의 비율을 FAR과 FRR의 그래프로 나타내었다.



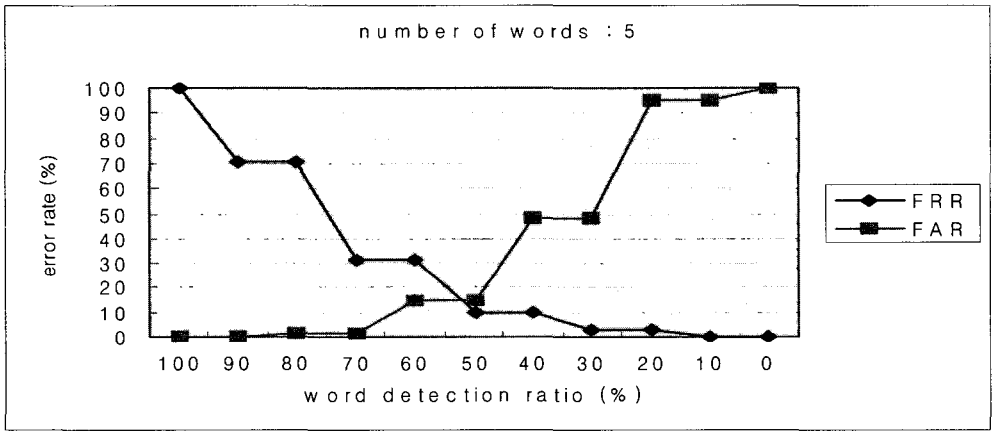

(a) 2단어



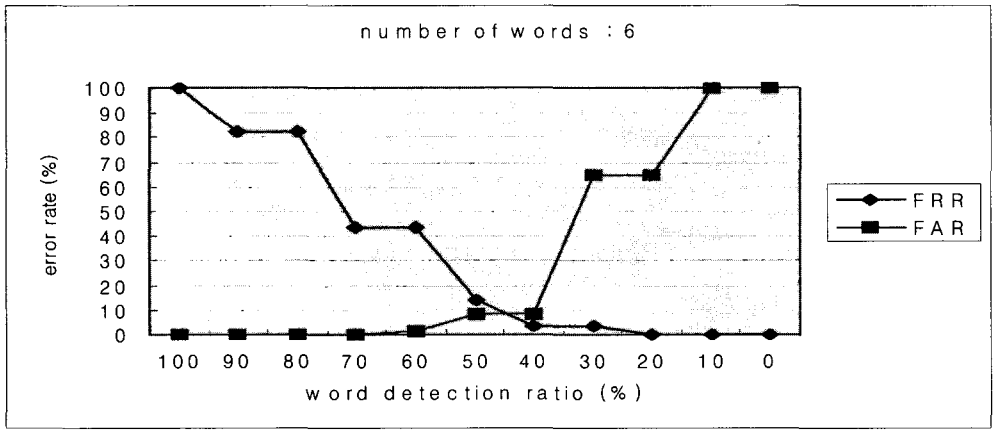

(b) 3단어




(c) 4단어



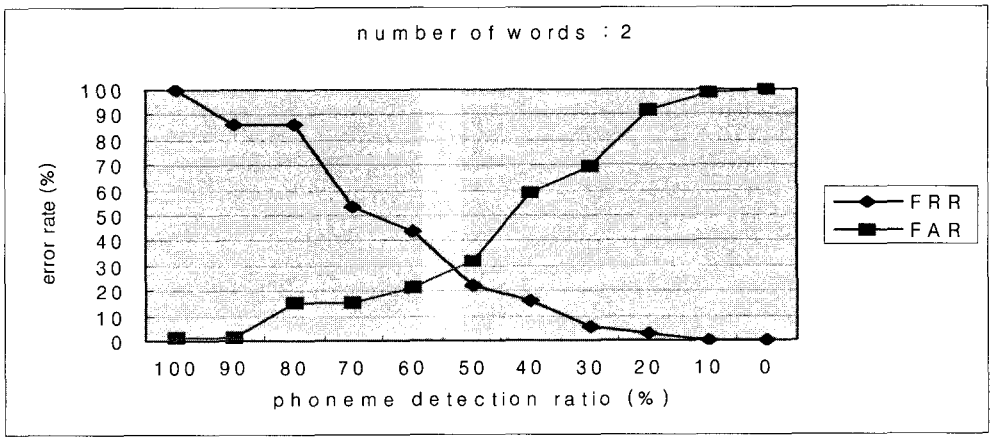

(d) 5단어



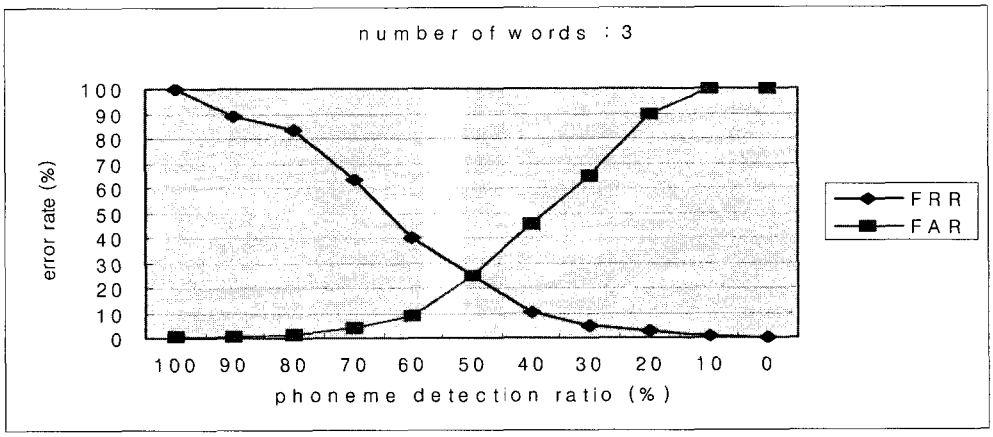

(e) 6단어

<그림 12> 문장의 단어 수준 검출률 인식 네트워크에 따른 결과 (방법 4)

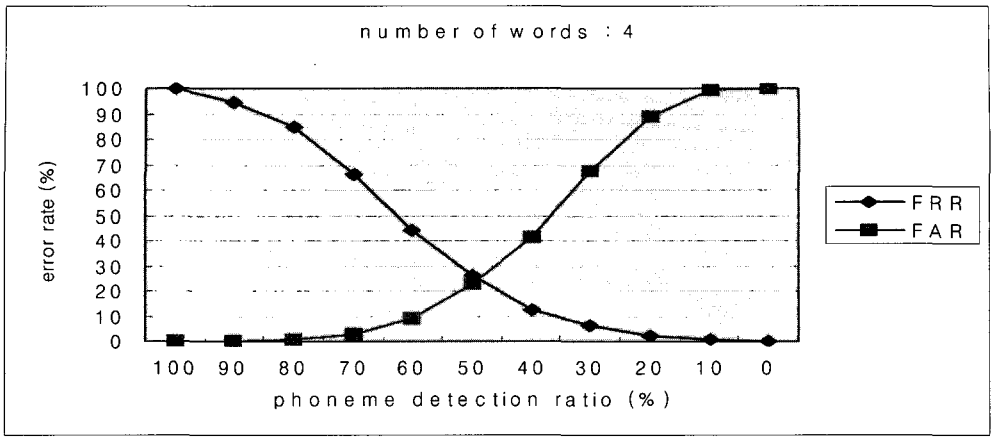
<그림 13>은 음소 수준의 검출률을 이용한 방법(방법 5)의 결과로서, 인식 네트워크를 통과한 문장의 인식 결과에 포함된 인식된 음소의 누적비율을 FAR과 FRR의 그래프로 나타내었다.



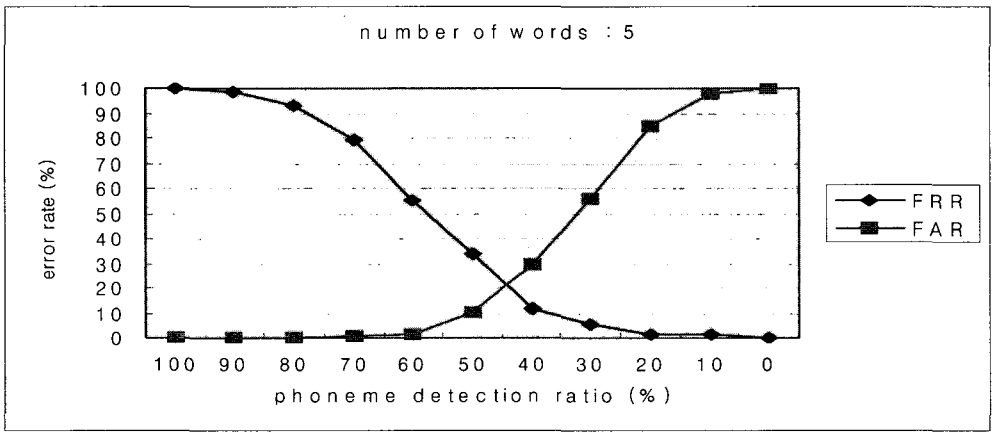

(a) 2단어



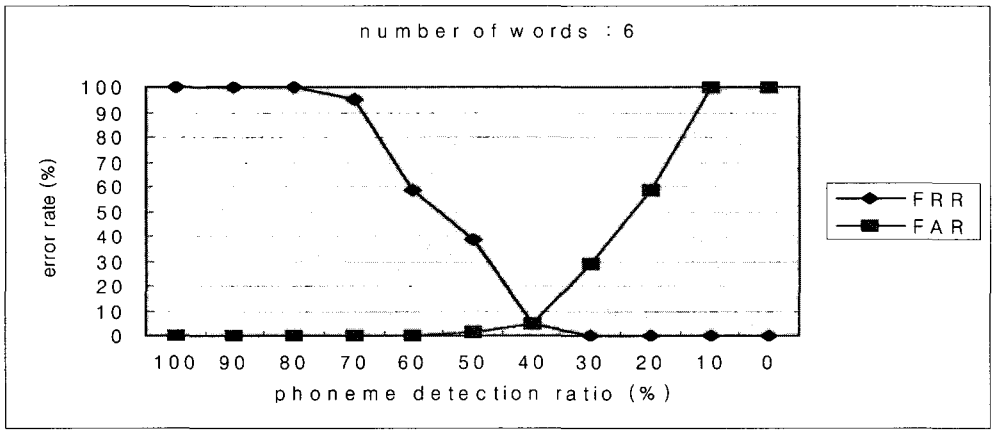

(b) 3단어




(c) 4단어




(d) 5단어




(e) 6단어

<그림 13> 문장의 음소 수준 검출률 인식 네트워크에 따른 결과 (방법 5)

음소수준의 검출률을 이용한 방법(방법 5)의 실험은 같은 단어의 수로 구성 된 인식 대상 문장일지라도, 문장 내 사용된 단어가 다르므로 그에 따른 단어들의 음소수도 틀려지게 된다. 즉, "Don't miss the bus."와 "I am a boy."는 같은 4단어로 이루어진 문장이지만, 문장을 구성하는 단어 들이 서로 다르므로, 단어를 구성하고 있는 음소의 수 또한 틀려진다. 따라서 (방법 5)의 네트워크에서 결과값을 문장의 단어수에 따른 분류가 아닌 문장 전체의 음소 수에 따라 분류해야 비교할 수 있다. <표 3>은 각각 문장의 음소수를 4가지로 분류하여 대상 문장 내 인식되지 않은 음소수를 조사하여 그 비율을 나타낸 것으로 FAR과 FRR의 평균값을 음소수준의 검출률 방법의 결과(Result05)와 비교할 수 있다.

<표 3> 단어별 음소 수에 따른 인식 음소 검출 백분율 누적값 (Result05_1)

| 검출 누락 비율 (%) / 단어별 음소수 | 0 | 1 ~ 10 | 11 ~ 20 | 21 ~ 30 | 31 ~ 40 | 41 ~ 50 | 51 ~ 60 | 61 ~ 70 | 71 ~ 80 | 81 ~ 90 | 91 ~ 100 | 비 고 | 문장수 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5,6,7,8 | 100 | 85.89 | 85.89 | 63.64 | 42.03 | 28.06 | 12.18 | 6.16 | 3.66 | 1.02 | 0.14 | FRR | 680 |
| | 0.29 | 0.29 | 3.96 | 6.6 | 12.77 | 26.59 | 42.32 | 55.55 | 85.6 | 99.71 | 100 | FAR | |
| 9,10,11 | 100 | 91.87 | 83.1 | 61.86 | 42.36 | 25.09 | 13.71 | 5.99 | 2.49 | 0.72 | 0 | FRR | 1,800 |
| | 0.27 | 0.38 | 1.76 | 3.76 | 9.98 | 24.09 | 47.31 | 71.73 | 90.95 | 99.89 | 100 | FAR | |
| 12,13,14 | 100 | 97.84 | 89.93 | 77.27 | 50.17 | 32.26 | 9.55 | 4.64 | 0.98 | 0.57 | 0.16 | FRR | 1,200 |
| | 0 | 0 | 0 | 1 | 3 | 16.16 | 32.82 | 57.43 | 84.34 | 97.5 | 100 | FAR | |
| 15,16,17 | 100 | 98.87 | 94.78 | 77.28 | 52.97 | 25.44 | 8.85 | 2.72 | 1.36 | 0 | 0 | FRR | 440 |
| | 0 | 0 | 0 | 0 | 0.9 | 6.58 | 24.76 | 57.06 | 88.19 | 99.55 | 100 | FAR | |

인식 성능을 평가하기 위하여 문장 내 단어 수에 따른 최적의 적응 문턱값을 가변적으로 적용 하여 평가 하였다. 식 (5)를 적용하여 각각의 방법에서 문장의 단어 수에 따른 오류가 최소가 되는 FAR과 FRR의 평균값을 <표 4>, <표 5>와 같이 구하였다.

$$Average\ of\ FAR\ \text{and}\ FRR = \frac{\sum_{i=2}^{6} average(FAR_i, FRR_i) \times N_i}{Total\#\ of\ sentences} \tag{5}$$

$$where\ N_i = \#\ of\ sentences\ \ of\ \ length\ \ i$$

<표 4> 필러 모델 방법을 사용한 단어수별 오류가 최소가 되는 최적의 평균값

| 단어수 | Result01 | | Result02 | | Result03 | | 문장수 |
|---|---|---|---|---|---|---|---|
| | 문턱값 $(\hat{\theta}_f)$ | FAR과 FRR의 평균(%) | 문턱값 $(\hat{\theta}_f)$ | FAR과 FRR의 평균(%) | 문턱값 $(\hat{\theta}_f)$ | FAR과 FRR의 평균(%) | |
| 2 | 10 | 16.25 | 60 | 18.43 | 70 | 19.69 | 160 |
| 3 | 10 | 17.02 | 70 | 15.96 | 10 | 16.83 | 1,280 |
| 4 | 40 | 12.03 | 70 | 21.56 | 10 | 22.29 | 1,640 |
| 5 | 50 | 10.31 | 70 | 20.45 | 10 | 22.65 | 960 |
| 6 | 60 | 9.38 | 70 | 10.63 | 10 | 21.25 | 80 |
| 평균 | | 13.29 | | 19.28 | | 20.56 | 4,120 |

<표 5> 검출률 방법을 사용한 오류가 최소가 되는 최적의 평균값

| 단어수 | Result04 | | Result05 | | 문장수 | 음소수 | Result05_1 | | 문장수 |
|---|---|---|---|---|---|---|---|---|---|
| | 문턱값 $(\hat{\theta}_d)$ | FAR과 FRR의 평균(%) | 문턱값 $(\hat{\theta}_d)$ | FAR과 FRR의 평균(%) | | | 문턱값 $(\hat{\theta}_d)$ | FAR과 FRR의 평균(%) | |
| 2 | 10 | 25.94 | 50 | 26.88 | 160 | 5,6,7,8 | 60 | 27.25 | 680 |
| 3 | 10 | 21.05 | 40 | 24.68 | 1,280 | 9,10,11 | 50 | 24.59 | 1,800 |
| 4 | 40 | 16.91 | 50 | 24.48 | 1640 | 12,13,14 | 60 | 21.19 | 1,200 |
| 5 | 50 | 12.23 | 60 | 20.865 | 960 | 15,16,17 | 50 | 16.01 | 440 |
| 6 | 60 | 6.25 | 60 | 5 | 80 | | | | |
| 평균 | | 17.25 | | 23.41 | 4,120 | 평 균 | | 23.12 | 4,120 |

<표 4>는 필러 모델을 사용한 각 방법의 결과이며, <표 5>는 검출률을 이용한 각각의 방법의 결과이다. 비인식 대상 문장 거부의 성능을 비교하여 보면 단어 수준 필러 모델을 사용한 방법(방법 1)이 13.29%로 가장 좋은 성능을 보였으며, 단어 수준의 검출률을 이용한 방법(방법 4)이 17.25%로 두 번째로 좋은 성능을 보였다. 따라서 문장 거부의 성능 면에서는 단어 수준의 필러 모델을 사용한 방법이 가장 좋은 성능을 나타낸 것을 알 수 있다. 또한 단어/음소 검출률을 이용한 방법 중 문장의 단어의 수에 따라 분류한 결과(Result05)와 음소의 수로 분류한 결과(Result05_1)를 비교하면 각각 23.41%과 23.12%로 크게 차이가 없었다. 모든 결과를 종합하여 보면 단어 수준의 방법이 음소 수준의 방법보다 더 나은 성능을 보

인다는 것을 알 수 있다.

또한 문장 거부 네트워크를 구성하는 방법에 따른 성능을 비교하여 보면, 필러 모델의 경우 성능 면에서는 우수하나, 입력된 모든 문장이 인식네트워크에서 각 단어별 표준 모델과 필러 모델들을 거쳐야 함으로 속도가 느려질 수 있다는 단점과, 인식 대상 마다 별도의 필러 모델을 만들어야 하므로 인식 대상의 수가 많거나 가변적인 경우 시스템 성능 저하를 가져올 수 있다. 반면 새로운 모델 추가가 필요 없고, 인식 네트워크의 모델들을 전부 거치지 않아도 되는 단어/음소 검출률을 이용하는 방법은 네트워크 구조상 모든 네트워크를 거치지 않아도 되므로 실행속도와 메모리 절약 면에서 효과적일 수 있다. 따라서 음성 인식 시스템의 사용 목적에 따라 이 두 방법을 달리 적용한다면 보다 큰 효과를 얻을 수 있다.

## 5. 결론 및 향후과제

본 논문에서는 음성 인식 시스템에서 거부 성능을 향상하는 방법을 통하여 음성 인식 시스템에서 사용자의 잘못된 발성 패턴이 입력되었을 때 인식대상 문장은 검출하되 비인식 대상 문장은 거부하여 정확도를 높이는 음성 시스템에 있어서 인식네트워크를 다양하게 구성함으로써 인식률과 거부율이 달라질 수 있음을 보였다.

본 실험은 인식 대상 문장을 거부하기 위해서 인식 대상 문장과 비인식 대상 문장 간에 음소기반의 필러 모델을 구성하였고, 단어/음소 수준의 검출률에 따른 문장 거부를 위하여 인식 대상 문장의 단어나 음소를 선택적으로 인식 할 수 있도록 모델을 구성하였으며, 각각의 방법들은 인식 네트워크를 거친 입력문장의 결과에 필러 모델의 비율과, 문장 내 인식된 단어나 음소의 비율의 포함률을 조사하는 방법을 택하였다. 그 결과 FAR과 FRR을 최소화 할 수 있는 최적의 평균값으로 각 방법들을 비교·분석할 수 있었는데, 두 방법에 있어서 음소수준보다 단어 수준의 모델에서 더 좋은 성능을 보였고, 단어 수준의 필러 모델의 경우 13.29%로서 가장 좋은 성능을 보였다. 또한 단어 수준의 검출률 모델은 17.25%로 단어 수준의 필러 모델을 사용한 방법보다 성능 면에서는 다소 떨어지지만, 인식 네트워크의 복잡도의 차이로 수행 속도와 메모리 절약 면에서 보다 효과적일 수 있어서 음성 인식 시스템의 목적에 따라 달리 적용한다면 큰 효과를 얻을 수 있다.

향후 연구에서는 문장 거부 성능을 높이는 방법으로 보다 다양한 음성 인식 네트워크 구성의 연구가 필요하며, 논문에서 제시한 단어별 필러 모델 방법과 단어 검출률 모델을 통합한 새로운 방법의 연구가 필요하다.

## 참고문헌

[1] 김무중 외, "한국인을 위한 영어 발음 교정 시스템의 개발 및 성능 평가", *말소리*, 제 46호, pp. 87-102, 2003.

[2] 김동화, 김형순, 김영호, "고립단어 인식 시스템에서의 거절기능 구현", *한국음향학회지*, 제 16권 제 6호, pp. 106-109, 1997.

[3] R. C. Rose, "Discriminant wordspotting techniques for rejection nonvocabulary utterances in unconstrained speech", *Proc. ICASSP*, Vol. 2, pp.105-108, Mar. 1992.

[4] 김무중, 김병기, 하진영, "음소기반 인식 네트워크에서의 비인식 대상 단어 거부 기능 성능 분석", *한국음향학회 하계학술발표논문집*, 제22권, No. 1(s), pp. 85-88, 2003.

[5] 이병혁, 하진영, "비인식 대상 문장 거부 기능을 위한 음소 기반 인식 네트워크의 구성에 관한 연구", *한국정보과학회 추계학술발표논문집*, Vol. 31, No. 2 pp. 772-774, 2004.

[6] 김우성, 구명완, "반음소 모델링을 이용한 거절 기능에 대한 연구", *한국음향학회지*, 제 18권, 제 3호, pp. 3-9, 1999.

[7] N. Moreau, D. Charlet and D. Jouvet, "Confidence measure and incremental adaptation for the rejection of incorrect data", *Proc. ICASSP*, pp. 1807-1810, 2000.

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

[9] H.-R. Kim, S. Yi and H.-S. Lee, "Out-of-variable vocabulary word recognition", *Proc. ICSP*, Vol, 1, pp. 337-339, 1999.

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

[11] B.-H. Lee and J.-Y. Ha, "Length dependent threshold for non-target sentence rejection", *The 4th Asia Pacific International Symposium on Information Technology*, pp. 62-465, Gold Coast, Australia, 2005.

[12] 박미나, 하진영, "HMM 모델링을 위한 HMM의 state수와 mixture수 분석", *한국정보과학회 봄학술발표논문집*, Vol. 29, No. 1, pp. 658-660, 2002.

[13] 구명완, 박상규, "HMM 음성 인식 시스템에서의 거절기능구현", *한국통신학회 학술발표회논문집*, 제 15권, 1호, pp. 94-97, 1996.

접수일자: 2006년 8월 9일
게재결정: 2006년 9월 16일

▸ 김형태(Hyung-Tai Kim)
주소: 200-701 강원도 춘천시 효자2동 192-1 강원대학교
소속: 강원대학교 컴퓨터정보통신공학부
전화: 011-9980-5258
E-mail: ds2swd@kangwon.ac.kr

▶ 하진영(Jin-Young Ha) : 교신저자
주소: 200-701 강원도 춘천시 효자2동 192-1 강원대학교
소속: 강원대학교 컴퓨터학부
전화: 033) 250-6386
E-mail: jyha@kangwon.ac.kr