가변어휘 단어 인식에서의 미등록어 거절 알고리즘 성능 비교

Performance Comparison of Out-Of-Vocabulary Word Rejection Algorithms in Variable Vocabulary Word Recognition

김 기 태*, 문 광 식*, 김 회 린**, 이 영 직***, 정 재 호*
(Ki-Tae Kim*, Kwang-Sik Moon*, Hoi-Rin Kim**, Young-Jik Lee***, Jae-Ho Chung*)

*인하대학교 전자공학과 디지털 신호처리 연구실, **한국정보통신대학원대학교 공학부,
***한국전자통신연구원 통신단말연구부 멀티모달 I/F팀

(접수일자: 1999년 10월 20일; 수정일자: 2000년 12월 1일; 채택일자: 2001년 1월 19일)

발화 검증이란 등록된 단어 목록 이외의 단어가 입력되었을 때, 미등록된 단어는 인식할 수 없는 단어임을 알려주는 기능으로써 사용자에게 친숙한 음성 인식 시스템을 설계하는데 중요한 기술이다. 본 논문에서는 가변어휘 단어 인식기에서 최소 검증 오류를 나타낼 수 있는 발화 검증 시스템의 알고리즘을 제안한다. 우선, 한국전자통신연구원의 PBW (Phonetically Balanced Words) 445DB를 이용하여 가변어휘 단어 인식에서의 미등록어 거절 성능을 향상시키는 효과적인 발화 검증 방법을 제안하였다. 구체적으로 특별한 훈련 과정이 없이도 유사 음소 집합을 많이 포함시킨 반음소 모델을 제안하여 최소 검증 오류를 지니도록 하였다. 또한, 음소 단위의 null hypothesis와 alternate hypothesis의 비를 이용한 음소 단위의 신뢰도는 null hypothesis로 정규화해서 강인한 발화 검증 성능을 보여 주었으며, 음소 단위의 신뢰도를 이용한 단어 단위의 신뢰도는 등록어와 미등록어 사이의 분별력을 잘 표현해 주었다. 이와 같이 새로이 제안된 반음소 모델과 발화 검증 방법을 사용했을 때, CA (Correctly Accept for Keyword: 등록어를 제대로 인정한 경우)는 약 89%, CR (Correctly Reject for OOV (Out-of-Vocabulary): 미등록어에 대해 거절한 경우)은 약 90%로써, 기존 필러 모델을 이용한 방법보다 미등록어 거절 성능이 ERR (Error Reduction Rate) 측면에서 약 15~21% 향상됨을 알 수 있었다.
핵심용어: 미등록어 거절
투고분야: 음성처리 분야 (2.5)

Utterance verification is used in variable vocabulary word recognition to reject the word that does not belong to in-vocabulary word or does not belong to correctly recognized word. Utterance verification is an important technology to design a user-friendly speech recognition system. We propose a new utterance verification algorithm for no-training utterance verification system based on the minimum verification error. First, using PBW (Phonetically Balanced Words) DB (445 words), we create no-training anti-phoneme models which include many PLUs(Phoneme Like Units), so anti-phoneme models have the minimum verification error. Then, for OOV (Out-Of-Vocabulary) rejection, the phoneme-based confidence measure which uses the likelihood between phoneme model (null hypothesis) and anti-phoneme model (alternative hypothesis) is normalized by null hypothesis, so the phoneme-based confidence measure tends to be more robust to OOV rejection. And, the word-based confidence measure which uses the phoneme-based confidence measure has been shown to provide improved detection of near-misses in speech recognition as well as better discrimination between in-vocabularys and OOVs. Using our proposed anti-model and confidence measure, we achieve significant performance improvement; CA (Correctly Accept for In-Vocabulary) is about 89%, and CR (Correctly Reject for OOV) is about 90%, improving about 15~21% in ERR (Error Reduction Rate).
Key words: *Out-of-vocabulary word rejection*
Subject classification: *Speech signal processing* (2.5)

책임저자: 김기태 (ktkim91@lgic.co.kr)
153-023 서울시 금천구 가산동 60-39
LG전자 미디어통신실
(전화: 02-818-4920; 팩스: 02-818-4802)

I. 서 론

대부분의 음성 인식 시스템은 사용자가 음성을 이용하여

사용할 수 있기 때문에 매우 자연스럽고 편리한 인터페이스를 제공한다. 그러나, 인식기에 등록이 안된 음성을 발성하면 이를 처리할 수 없다는 단점을 지니게 되므로 사용자는 정해진 등록어만을 사용해야 하는 제약을 받는다. 이러한 문제를 해결하기 위해서 미등록어 거절 (Out-Of-Vocabulary rejection)기능이 연구되어 왔는데, 이는 인식 대상 단어에 대해서만 인식을 하고 그 외에는 인식 결과를 내지 않고 거절함으로써 시스템의 성능을 향상시키고자 하는 것이 목적이다.

인식 거절은 구현 방식에 따라서 핵심어 검출 (keyword spotting) 방식과 발화 검증 (utterance verification) 방식으로 구분된다. 핵심어 검출 방식이란 문법을 설계할 경우 핵심어만 고려하고 그 이외의 단어는 garbage 모델을 사용하여 필요없는 단어를 제거하는 방식이다. 제거하는 방법은 garbage 모델의 유사도 값이 인식 대상 핵심어의 유사도 값보다 클 경우이다. 발화 검증 방식이란 인식결과를 확인하는 과정이 추가되며 이 때 필러 (filler) 모델을 이용하는 방법이 사용되었다. 그러나 이러한 필러 모델은 단어를 기반으로 구성되었기 때문에 가변 어휘 단어 인식 시스템을 위한 발화 검증이 구현되기 위해서는 매 음소 단위의 검증 기능이 필요하다. 이를 위해서 반음소 모델 (anti-phoneme model)을 사용하는 방식이 제안되었다[1].

본 논문에서는 성능이 우수한 발화 검증 방식을 제안했으며, 발화 검증의 역할은 가변어휘 단어 인식기에서 인식된 단어가 등록어인지 미등록어인지를 판별하는 것이다. 음성 인식의 많은 응용분야에서 통계적인 hypothesis 테스트를 이용하여 핵심어 검출과 발화 검증이 수행된다. 일반적으로 유사도 비를 이용한 테스트를 많이 사용하는데, 그 방법은 입력 단어가 등록어라고 가정하는 null hypothesis와 미등록어라고 가정하는 alternative hypothesis와의 비를 이용하는 것이다[1][2][3][4]. Alternative hypothesis는 2가지의 카테고리를 포함하고 있다. 즉, 미등록어와 잘못 인식된 등록어가 이에 해당된다. 그러나 잘못 인식된 등록어는 아주 미소하므로, 본 논문에서는 2가지의 카테고리중 미등록어 거절 기능에 초점을 맞추었다.

본 논문에서 가변어휘 단어 인식기는 비터비 탐색을 하므로 기본적으로 단어 단위로 인식이 되지만, 그 인식된 단어는 내부적으로 음소 단위로 인식이 된다. 따라서 최소 검증 오류를 갖는 반음소 모델을 제안하고, 이를 이용하여 인식된 음소 단위들을 각각의 반음소 모델과 비교하여 신뢰도를 구한다. 이 음소 단위의 신뢰도를 단어 단위의 신뢰도로 환산하기 위해서 음소 단위의 신뢰도를 평균 내는 방식을 취한다. 이렇게 함으로써, 등록어와 미등록어 사이의 분별력을 크게 하였다.

II. 가변 어휘 단어 인식 시스템

2.1. 가변 어휘 단어 인식 시스템의 개요[5]

가변 어휘 단어 인식기는 그림 1에서와 같은 구조를 갖는다. 이 인식기는 기존의 인식기와는 달리 인식 대상으로 하는 단어 목록이 매 음성 입력마다 바뀌어도 인식할 어휘에 대한 음성 훈련 과정을 새로 수행하지 않고 단지 발음 사전만을 새로 교체하여 단어 모델들을 재구성하므로 이론적으로 무제한의 임의의 단어를 주어진 단어 목록 내에서 인식할 수 있게 된다. 이러한 단어 인식기를 구현하려면, 우선 한국어에 존재하는 모든 음소를 충분한 음소 환경에서 정확히 모델링해야 한다. 이렇게 하기 위해서는 먼저 각 음소를 정확히 모델링하기 위하여 훈련 데이터를 다양한 음소 환경하에서 수집해야 하며, 또 이를 음소 모델에 적절히 반영시키기 위하여 이러한 다양성을 포용할 수 있는 음소 모델 구조를 가져야 한다. 이러한 조건을 충족시키기 위하여 본 연구에서는 훈련용 음성 데이터로써 ETRI가 보유하고 있는 PBW (Phonetically Balanced Words) 445DB를 사용하고, 39개의 문맥 독립 음소 모델을 사용하였다.

마지막으로, 매 어휘 변경 요구가 사용자로부터 입력되면 이를 즉시 단어 목록에 반영하고 이로부터 각 단어의 발음 사전이 사전 생성기를 통하여 출력될 수 있도록 하는 기능이 필요하다. 이 기능도 ETRI가 보유하고 있는 발음 사전 생성기를 이용하여 구현하였다.

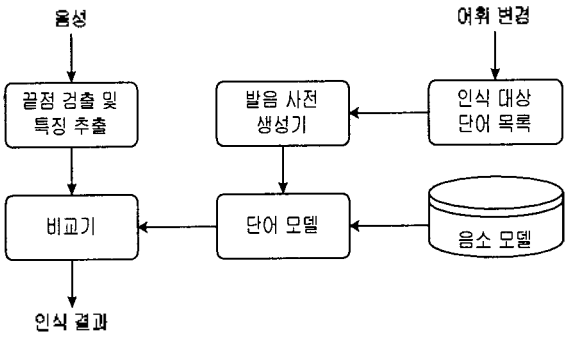

그림 1. 가변 어휘 단어 인식 시스템 구조
Fig. 1. Block diagram of variable vocabulary word recognizer.

2.2. 문맥 독립 음소 모델 훈련

가변 어휘 음성 인식기의 훈련 및 성능 분석에 사용된 음성 데이터베이스는 PBW 445 DB 이다.

2.2.1. PBW 445DB

다양한 음소의 조합을 고려한 PBW 445DB는 어휘수가 총 445개로 구성되어 있으며, 이를 1명이 2회 발성한 것을 1개의 set으로 하였다. 이러한 set이 남성음에 대하여 22set, 여성음에 대해서는 19set이 있어서 모두 합하면 41set (36,490개 단어)이 된다. 이 41set 중 30set를 훈련용 데이터로 선정하고, 이 중 356개의 단어를 문맥 독립 음소 모델 훈련에 사용하였다.

테스트용 데이터와 평가용 데이터는 각각 6set와 5set를 사용하였으며, 이 set는 훈련용 데이터 set와 중복되지 않게 선정하였다. 또한, 테스트용 데이터와 평가용 데이터는 등록어 데이터 45개 단어와 미등록어 데이터 44개 단어로 구분하고, 훈련용 데이터의 356개의 단어와 중복

되지 않게 하였다. 테스트용 데이터란 미등록어 거절 기능 알고리즘에서 임계값 (threshold)을 결정할 때 사용하는 데이터이며, 평가용 데이터는 말 그대로 객관적인 평가만을 위한 데이터이다. PBW 445DB의 A/D 방식은 16KHz, 16Bit이다.

2.2.2. 특징 추출 및 훈련

본 시스템은 통신환경에서의 적용을 고려하여 16KHz, 16Bit인 PBW 445DB를 8KHz, 16Bit로 변환하였다. 변환된 음성 샘플에 1차 차분 방정식

$$s_n' = s_n - ks_{n-1} \quad (1)$$

을 적용하여 pre-emphasis를 취했으며, pre-emphasis 계수 k는 0.97로 하였다. 특징 벡터를 추출하기 위해서 다음과 같은 조건들을 두었다. Hamming window를 취했으며, 창 크기 (window size)는 32msec (256sample)로 하였고, 창의 overlap 크기는 22msec로 하였다. 즉, 10msec (80 sample)씩 진행하면서 특징 벡터를 추출하였다. 특징 벡터로는 MFCC (Mel Frequency Cepstral Coefficient) 12차를 사용하였으며, 필터 뱅크의 채널 수는 12개를 사용하였다. 이 때 cepstral liftering은 수식

$$c_n' = \left(1 + \frac{L}{2}\sin\frac{\pi n}{L}\right)c_n \quad (2)$$

을 사용했고, cepstral liftering 계수 L은 22로 하였다. 또한 추가적으로 cepstral C_0와 델타 계수 (delta coefficient)를 사용하였다. 델타 계수에 대한 수식은

$$d_t = \frac{\sum_{\theta=1}^{\Theta}\theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta}\theta^2} \quad (3)$$

이며, delta window size Θ는 2로 하였다. 또한, CMS (Cepstral Mean Subtraction)를 취해서 채널 보상 효과를 고려했으며, 이와 같은 일련의 과정을 거쳐서 최종적으로 26차 특징 벡터를 구했다.

정의된 문맥 독립형 음소 39개 모델 (묵음 모델 포함)의 훈련은 앞서 기술한 바와 같이 PBW 445DB 중 30set의 훈련용 데이터의 356개의 단어를 가지고 수행한다. 각 음소 (묵음 제외)는 CHMM (Continuous Hidden Markov Model)으로 모델링되며, 모델의 구조는 3-상태 left-to-right (no skip path) 모델, 8가지 (mixture)로 정의하였고, 묵음은 1-상태 모델, 16가지로 정의하였다. 또한 훈련시에 각 단어의 전후에 묵음 모델을 사용하였으며, PBW 445DB는 음소별로 레이블링이 되어 있지 않으므로 각 단어의 음성과 음소별 트랜스크립션 (transcription)만을 가지고 훈련하였다. 이 때에 forward-backward re-estimation (Baum-Welch re-estimation)을 이용하여 음소별로 훈련시켰다. 이와 같은 방법으로 최종적인 음소 모델을 얻어낼 수 있었다.

III. 제안된 발화 검증 시스템을 탑재한 가변 어휘 단어 인식 시스템

3.1. 기본 시스템 구성

기본 시스템 구성을 위하여 2단계 구조를 사용하였다. 2단계 구조란 인식기의 후처리 방식으로 검증 기능을 구현하는 방법이다. 그러므로 2단계 구조는 기존에 구현되어 있던 시스템을 크게 수정하지 않고 추가로 검증 과정만을 구현하여 사용하기 때문에 구현에 소요되는 시간을 단축시킬 수 있는 장점이 있다[1].

발화 검증 시스템을 설계할 때에는 세가지 문제를 해결해야 한다[4]. 첫째, 미등록어와 잘못 인식된 단어를 잘 선별할 수 있는 검증 모델 set에 기반한 적절한 신뢰도 (confidence measure)를 정의해야 한다. 둘째, 훈련 데이터에서 검증 오류를 최소화할 수 있도록 검증 모델을 적응시키는 훈련 과정을 선택해야 한다. 셋째, 유사도의 변화, 검증 임계값의 변화, 훈련과 테스트 상태의 변화에 강해야 한다. 본 논문에서 사용한 반음소 모델과 신뢰도는 위의 요건들을 최대한 만족시키고자 하였다.

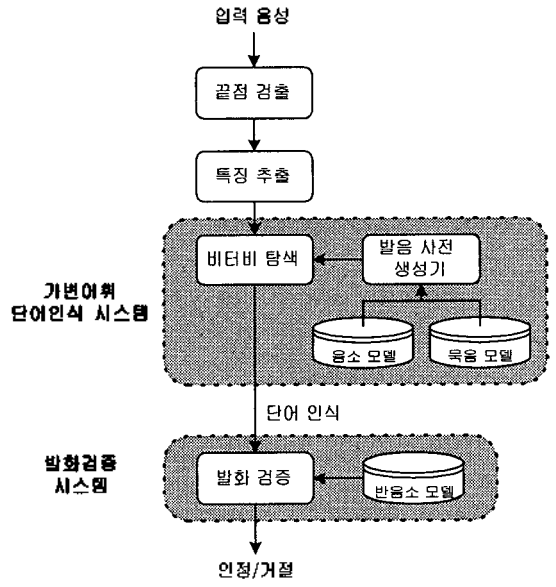

그림 2. 발화 검증 시스템을 탑재한 가변 어휘 단어 인식 시스템
Fig. 2. Variable vocabulary word recognizer with utterance verification system.

미등록어 거절 기능을 갖는 가변 어휘 단어 인식 시스템의 구성도는 그림 2에서 보는 바와 같이 가변 어휘 단어 인식 시스템 뒤에 발화 검증 시스템이 따라 붙는 2단계 시스템이다. 먼저 음성이 입력되면 끝점 검출기에 의해 음성 구간만 검출되며, 검출된 음성은 특징 추출 과정을 거치게 된다. 다음 첫번째 단계 (가변 어휘 단어 인식 시

스템)에서 39개의 음소 모델들을 사용해서 비터비 탐색 알고리즘에 의한 인식 과정이 수행된다. 음소 모델들은 ML (Maximum Likelihood) criteria를 이용하여 HMM의 파라미터를 최적화시켰다. 인식 과정 동안 각 단어의 발화는 음소 hypothesis로 segmentation되며, 그 결과를 발화 검증 시스템으로 넘긴다. 두번째 단계 (발화 검증 시스템)에서 검증 과정은 인식된 후보 단어의 음소열에 대해 반음소 모델과의 신뢰도를 구해 그 단어의 신뢰도 값을 결정한다. 이 신뢰도 값이 미리 정해둔 임계값보다 크면 그 인식 단어로 인식하고 아니면 거절한다[1][4].

3.2. 기존 발화 검증 방법[6]

음소 필러 모델을 사용한 핵심어 검출 방식 (Keyword Spotting Method)을 이용해서 미등록어를 거절시키는 방법이다. 핵심어 검출 방식은 핵심어 (등록어)모델과 필러 모델 (garbage 모델)을 사용하는 연결단어 인식 알고리즘을 기반으로 하고 있다. 여기서 필러 모델들은 핵심어에 해당하지 않는 음성 구간들, 즉 비핵심어들과 비음성, 즉 묵음 또는 배경 잡음 구간들을 표현하는데 사용된다.

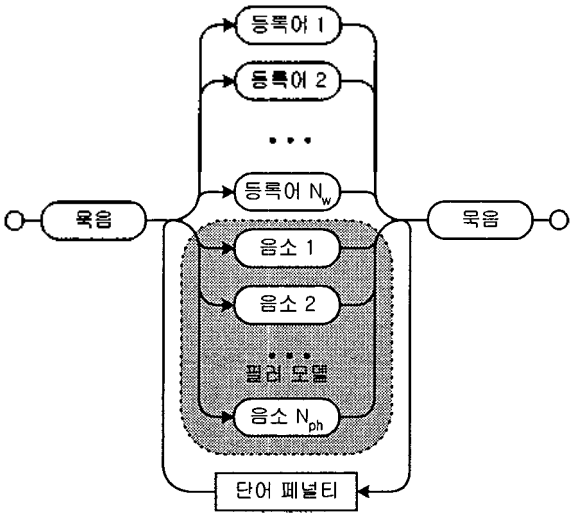

그림 3. 기존 가변 어휘 단어 인식 시스템의 네트워크
Fig. 3. The network of existing variable vocabulary word recognizer.

기존 미등록어 거절 방법에서는, 가변 어휘 단어 인식 시스템에서 비터비 탐색시 사용되는 네트워크 망을 그림 3과 같이 구성한다. 구성된 네트워크 망에서 인식된 결과는 등록어들과 음소들의 열로 나타나게 된다. 즉, "묵음 + (등록어 및 음소들의 열) + 묵음"과 같은 형태가 된다. 여기에서 단어 페널티를 잘 조정하면, 입력된 음성이 등록어이면 인식된 결과는 "묵음 + (등록어 및 약간의 음소들의 열) + 묵음"으로 나타나게 되고, 미등록어이면 인식된 결과는 "묵음 + (등록어 및 다수의 음소들의 열) + 묵음" 또는 "묵음 + (다수의 음소들의 열) + 묵음"으로 나타나게 된다. 이 인식된 결과를 발화 검증 시스템으로 넘긴다. 발화 검증 시스템에서는, 가변 어휘 단어 인식 시스템의 단어 페널티와 인식된 결과의 삽입된 음소들의 개수를 이용하여 미등록어를 거절시킬 수 있다. 삽입된 음소들은 필러 모델을 뜻하며, 삽입된 음소가 많다는 것은 인식 결과에 핵심어가 없다는 의미이다. 즉 사용자가 미등록어를 발성하게 되면, 필러 모델들로 인식됨을 알 수 있다. 또한, 삽입된 음소가 미리 정해둔 임계값 이하라도 인식 결과에 등록어가 포함되어 있지 않거나 2개 이상이면, 무조건 거절시킨다.

3.3. 제안한 발화 검증 방법

발화 검증 방식에서는 단어나 음소 단위의 인식 결과를 받아들일 것인지, 거절할 것인지를 결정하는 검증 과정이 이용된다. 본 논문에서는 반음소 모델을 사용한 발화 검증 방식을 이용해서 미등록어를 거절시켰으며, 39개의 잘 훈련된 음소 모델만 있으면, 반음소 모델을 만들기 위한 특별한 훈련을 거치지 않고 반음소 모델을 만들 수 있도록 제안하였다. 또한 음소 단위의 신뢰도를 잘 이용하여서, 가변 어휘 단어 인식 시스템에서 사용할 수 있는 단어 단위의 신뢰도를 구성해 보았다. 여기서 신뢰도란 음성 인식 결과에 대해서 그 결과가 얼마나 믿을만한 것인가를 나타내는 척도로서, HMM 모델의 비터비 탐색 결과 수치와는 다른 것이다. 비터비 탐색 결과 수치는 어떤 단어나 음소에 대한 단순한 유사도를 나타내지만, 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 유사도로도 볼 수 있다. 3.3.1절에서는 본 시스템에서 사용된 비터비 탐색시의 네트워크를 기술하고, 3.3.2와 3.3.3절에서는 제안한 반음소 모델을 구성하는 법과 단어 단위의 신뢰도를 구하는 방법을 소개한다.

3.3.1. 본 시스템에서 사용된 가변 어휘 단어 인식 시스템의 네트워크

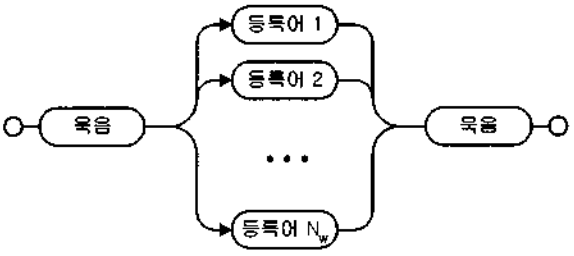

그림 4. 본 시스템에서 사용된 가변 어휘 단어 인식 시스템의 네트워크
Fig. 4. The network used in proposed system.

본 시스템에서 사용된 가변 어휘 단어 인식 시스템의 네트워크는 그림 4와 같다. 이와 같이 등록어들로만 구성된 네트워크를 사용하면, 인식되는 결과는 "묵음 + 등록어 + 묵음"과 같은 형태가 된다. 즉, 입력된 음성이 등록

어이든, 미등록어이든 가변 어휘 단어 인식 시스템에서는 항상 등록어로 인식을 하게 된다. 이 인식된 결과를 제안된 발화 검증 시스템으로 넘겨서 인식된 결과가 정말 등록어인지 아니면 미등록어인지를 판별한다.

3.3.2. 반음소 모델

반음소 모델은 자기 음소를 제외한 유사 음소 집합을 말한다. 일반적으로 유사 음소 집합이 많을수록 반음소가 잘 모델링되지만, 유사 음소 집합의 크기가 너무 크게 되면 훈련 데이터량이 너무 많아지는 단점이 있다[1].

본 논문에서 제안한 반음소 모델은 이 두가지를 모두 만족시키는 새로운 반음소 모델을 구현했다. 즉 39개의 잘 훈련된 음소 모델만 있으면, 반음소 모델을 만들기 위한 특별한 훈련을 거치지 않고 반음소 모델을 만들었으며[3], 또한 유사 음소 집합도 자기 음소와 묵음을 제외한 나머지 모든 음소를 모두 포함시켜서 반음소 모델량이 잘 되도록 하였다. 이렇게 함으로써 검증 오류의 최소화를 추구하였다. 구체적으로, 자기 음소를 제외한 나머지 음소 ("묵음" 제외)들의 best Gaussian, 2nd best Gaussian, 3rd best Gaussian, 4th best Gaussian의 가중값, 평균, 분산을 취한다. 따라서 각 음소의 반음소 모델은 3상태를 가지고 각 상태는 148개의 가지 (각 음소당 4가지 37개 음소 (39개 음소 자기음소 묵음음소))를 소유한다. 천이 확률은 자기 음소를 제외한 나머지 음소 ("묵음" 제외)들의 평균을 취한다. 한편, 5th best Gaussian이상을 포함하면, 미등록어 거절 성능은 다소 향상되지만 계산 속도가 저하된다.

3.3.3. 단어 단위의 신뢰도

본 시스템에서는 가변 어휘 단어 인식기를 이용하여 비터비 탐색을 하므로 기본적으로 단어 단위로 인식이 되지만, 그 인식된 단어는 내부적으로 음소 단위로 인식이 된다. 따라서 인식된 음소 단위들을 각각의 반음소 모델과 비교하여 신뢰도를 구하고, 음소 단위의 신뢰도를 단어 단위의 신뢰도로 환산하기 위해서 음소 단위의 신뢰도를 평균 내었다.

우선 38개의 다른 패턴, 즉 $\Theta=\{\Theta_1,\ldots,\Theta_i,\ldots,\Theta_{38}\}$에 상응하는 발화 검증 모델을 사용하는 신뢰도를 선택했다. 각 패턴 i에 대해서 음소 모델을 $\Theta_i^{(h)}$라 표시하고, anti-model인 반음소 모델을 $\Theta_i^{(a)}$라 표시했다 (즉, $\Theta_i=\{\Theta_i^{(h)},\Theta_i^{(a)}\}$). 따라서 음소 단위들을 평균낸 단어 단위의 신뢰도[4]는

$$s_i(\mathbf{O};\Theta)=\log\left[\frac{1}{N(i)}\sum_{q=1}^{N(i)}\exp\left\{f\cdot Lr_{i(q)}(\mathbf{O}_q;\Theta)\right\}\right]^{\frac{1}{f}} \quad (4)$$

와 같이 되며, 이 신뢰도가 미리 정해둔 임계값 τ_s이하라면 거절시키게 된다. 여기서 f는 음의 값을 가지는 상수이며, 가변 어휘 단어 인식기에서 인식된 결과인 등록어 i는 $M(i)$음소들로 구성되어 있다. 각 음소의 반음소 모델과의 유사도 비 거리, $Lr_{i(q)}(O_q;\Theta)$[2]는

$$Lr_{i(q)}(O_q;\Theta)=\frac{g_{i(q)}(O_q)-G_{i(q)k(O_q)}}{|\,g_{i(q)}(O_q)\,|} \quad (5)$$

와 같이 정의하였다. 수식 (5)는 개선된 음소 단위의 신뢰도로도 말하여질 수 있으며, log-유사도 값의 크기, $|\,g_{i(q)}(O_q)\,|$으로 정규화 함으로써 프레임 길이로 정규화 하는 것보다 음성 인식에 사용되는 문법의 변화에 탄력이 있으며, 일관적인 음소 단위의 검증 성능을 나타낸다. 패턴 l (즉, $l=i(q)$)인 일반적인 음소에 대해서

$$g_{i(q)}(O_q)=\log p(O_q\mid\Theta_l^{(h)}) \quad (6)$$

$$G_{i(q)}(O_q)=\log p(O_q\mid\Theta_l^{(a)}) \quad (7)$$

이 적용되며, 수식 (6)과 (7)을 계산할 때 사용되는 관측 확률은

$$b_j(\mathbf{o})=\max_{1\le k\le(8\text{ or }148)}\left\{c_{jk}N(\mathbf{o},\mu_{jk},\mathbf{U}_{jk})\right\} \quad (8)$$

과 같이 각 상태에서 최대 스코어를 내는 가지만을 이용하였다. 일반적으로 관측 확률을 계산할 때에는 $\sum$를 사용하여 모든 가지의 스코어를 더하지만, 본 논문에서는 **max** 를 사용하여 최대 스코어를 내는 가지만을 사용하는 것이 미등록어 거절 성능을 향상시킨다는 것을 실험을 통하여 확인하였다.

여기서, c_{jk}는 가지의 가중값이며 j는 각 음소의 상태를 뜻하며, k는 각 상태의 가지를 뜻한다. N은 각 가지의 Gaussian 분포를 의미하며, u_{jk}는 평균 벡터, U_{jk}는 covariance matrix이다.

수식 (4)와 같은 신뢰도는 음성 인식에서 근소한 오류의 검출을 잘해줄 뿐만 아니라, 등록어와 미등록어 사이의 분별력을 더 잘 보여준다. 실제로 등록어를 발화 검증 해보면 $s_i(O;\Theta)$가 발화 검증 임계값 τ_s보다 크며, 미등록어를 발화 검증 해보면 τ_s보다 작다.

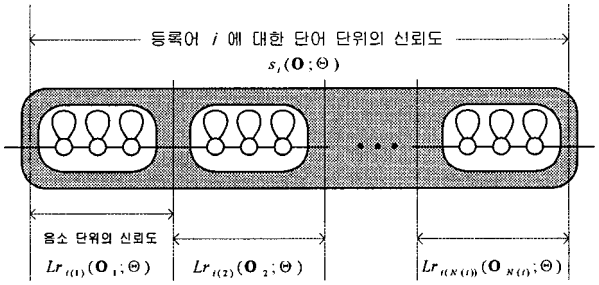

그림 5. 음소 단위의 신뢰도 및 단어 단위의 신뢰도
Fig. 5. Phoneme-based confidence measure and word-based confidence measure.

IV. 실험 결과

4.1. 성능 평가 기준(1)

미등록어 거절의 성능 평가는 다음의 항목을 기준으로 삼았다.

● 등록어

(1) CA : Correctly Accepted for Keyword, 즉 인식 대상 등록어를 제대로 accept한 경우의 확률

(2) FAI : False Accepted In-Grammar Word (=Keyword), 즉 인식 대상 등록어로 accept는 했지만 잘못 인식한 경우의 확률

(3) FR : False Rejected for Keyword, 즉 인식 대상 등록어를 말했는데 reject한 경우의 확률

(4) 따라서 CA + FAI + FR = 100% 이다.

● 미등록어

(1) CR : Correctly Rejected for OOV, 즉 미등록어에 대해 reject한 경우의 확률

(2) FAO : False Accepted Out-of-Grammar Word(=OOV), 즉 미등록어인데 accept한 경우의 확률

(3) 따라서 CR + FAO = 100% 이다.

계산 속도는 성능 평가 기준에 포함시키지 않았다. 제안한 발화 검증 방법은 실제 구현시 최적화할 수 있는 여지가 많기 때문이다.

4.2. 기존 발화 검증 방법의 실험

기존의 미등록어 거절 방법은 3.2절에서 자세히 소개하였다. 표 1과 2는 단어 페널티와 삽입된 음소들의 개수에 따른 성능을 보여준다. 이 방법은 단어 페널티와 인식된 결과의 삽입된 음소 개수를 이용하여 미등록어를 거절시키므로, 임계값의 요소는 단어 페널티와 삽입된 음소의 개수, 즉 2개가 된다. 시험용 데이터에서 CA와 CR이 같은 경우를 표 1에서 회색으로 표시하였다. 그러나, 같은 임계값에 대해서 평가용 데이터는 CA와 CR이 같지 않음으로써 일관적인 미등록어 거절 성능을 보이지 않았다.

기존 미등록어 거절 방법은 원하는 성능을 발휘하기 위해서는 2개의 요소를 같이 조정해야 하므로 다소 힘든 작업이 되고, 미등록어 거절 성능도 제안된 발화 검증 시스템보다 떨어진다.

표 1. 기존 발화 검증 방법의 성능 (단어 페널티 l=-30.0일 때)
Table 1. Performance of existing utterance verification method (word penalty l=-30.0).

데이터 종류 / 삽입된 음소개수	시험용 데이터					평가용 데이터				
	CA	CR	FAI	FAO	FR	CA	CR	FAI	FAO	FR
4개이상이면, 거절	85.37	69.13	4.44	30.87	10.19	89.14	69.19	3.70	30.81	7.16
3개이상이면, 거절	84.81	76.89	3.70	23.11	11.49	88.89	72.98	3.45	27.02	7.66
2개이상이면, 거절	83.52	83.71	2.96	16.29	13.52	88.15	82.83	2.46	17.17	9.39
1개이상이면, 거절	66.30	94.70	1.48	5.30	32.22	73.09	92.17	1.23	7.83	25.68

표 2. 기존 발화 검증 방법의 성능 (단어 페널티 l=-40.0일 때)
Table 2. Performance of existing utterance verification method (word penalty l=-40.0).

데이터 종류 / 삽입된 음소개수	시험용 데이터					평가용 데이터				
	CA	CR	FAI	FAO	FR	CA	CR	FAI	FAO	FR
4개이상이면, 거절	90.37	49.43	4.26	50.57	5.37	94.07	52.27	3.21	47.73	2.72
3개이상이면, 거절	90.00	54.36	3.89	45.64	6.11	94.07	56.06	3.21	43.94	2.72
2개이상이면, 거절	89.44	69.32	3.70	30.68	6.85	93.82	67.42	2.47	32.58	3.71
1개이상이면, 거절	76.67	87.88	2.04	12.12	21.29	82.22	84.34	1.73	15.66	16.05

4.3. 본 논문에서 제안한 발화 검증 방법의 실험

본 논문에서 제안한 발화 검증 방법은 3.3절에서 자세히 소개하였다. 여기서는 임계값 별 발화 검증 시스템의 성능을 보았으며, 임계값은 음소 단위의 신뢰도를 평균낸 단어 단위의 신뢰도, 즉 수식 (4)를 이용한 값이다.

표 3은 반음소 모델 구현시 best Gaussian, 2nd best Gaussian까지 포함한 결과이며, 표 4는 best Gaussian, 2nd best Gaussian, 3rd best Gaussian, 4th best Gaussian 까지 포함한 결과이다. 여기에서 알 수 있듯이, 반음소 모델은 자기 음소를 제외한 다른 모든 음소의 다양한 Gaussian을 포함할수록 미등록어 거절 기능의 성능이 향상되었으며, 또한 강인한 알고리즘이 됨을 확인할 수 있었다. 구체적으로, 표 3에서 시험용 데이터는 임계값 -0.063에서 CA와 CR이 같았으며 평가용 데이터는 임계값 -0.068에서 CA와 CR이 같음으로써, 서로 다른 임계값에서 CA와 CR이 같음을 보여 주었다. 즉, 시험용 데이터와 평가용 데이터 사이에 일관적인 성능을 보이지 못했다.

표 3. 본 논문에서 제안한 발화 검증 방법의 성능 (반음소 모델 구현시 best Gaussian, 2nd best Gaussian까지 포함)
Table 3. Performance of proposed utterance verification method (For anti-phoneme model including best Gaussian, 2nd best Gaussian).

데이터 종류 / 임계값	시험용 데이터					평가용 데이터				
	CA	CR	FAI	FAO	FR	CA	CR	FAI	FAO	FR
-0.060	84.63	87.50	1.48	12.50	13.89	83.95	91.67	0.99	8.33	15.06
-0.062	85.75	86.93	1.48	13.07	12.77	85.18	91.16	0.99	8.84	13.83
-0.063	86.12	86.17	1.48	13.83	12.40	86.91	90.66	0.99	9.34	12.10
-0.064	86.67	85.80	1.48	14.20	11.85	87.16	89.39	0.99	10.61	11.85
-0.066	87.04	84.66	1.66	15.34	11.30	87.90	88.64	0.99	11.36	11.11
-0.068	88.52	83.71	1.66	16.29	9.82	88.64	88.38	1.24	11.62	10.12
-0.070	89.45	83.71	1.66	16.29	8.89	90.12	87.37	1.24	12.63	8.64
-0.072	89.45	82.95	1.66	17.05	8.89	91.35	86.11	1.24	13.89	7.41

반면, 표 4에서는 임계값 -0.094에서 테스트 데이터와 평가용 데이터가 동시에 CA와 CR이 동일함을 보임으로써, 일관적이면서도 강력하게 미등록어 거절 성능을 보여주었다. 표 4에서 알 수 있듯이 시험용 데이터에 대해서는 CA는 87%, CR은 87%정도의 성능을 보여 주었으며, 평가용 데이터에 대해서는 CA는 89%, CR은 90%정도의 성능을 나타내었다. 따라서 best Gaussian, 2nd best Gaussian, 3rd best Gaussian, 4th best Gaussian까지 포함한 반음소 모델을 이용할 경우가 가장 성능이 우수함을 알 수 있으며, 임계값은 0.094에서 잡는 것이 가장 적당함을 알 수 있었다.

그림 6은 기존 발화 검증 방법과 제안한 발화 검증 방법에 대한 CA-CR 곡선이다. 이 곡선에서도 알 수 있듯이 제안한 발화 검증 방법이 기존 발화 검증 방법보다 CA-CR곡선이 우측 상단에 위치함으로써 CA-CR 성능이 보다 우수함을 확인할 수 있다.

표 5는 등록어와 미등록어를 구성하고 있는 음소 단위의 신뢰도 값을 나타낸다. 이 표에서 음소 단위의 신뢰도 값은 각 음소가 가지는 신뢰도 값들을 산술 평균낸 값이다. 표 4에서 임계값을 -0.094로 결정하였으므로, 등록어인 경우에는 음소 단위의 신뢰도 값이 -0.094 이하이면 오류가 있는 것이며, 미등록어인 경우에는 -0.093 이상이면 오류가 있는 것이며, 이탤릭 체로 진하게 표시하였다. 그러나, 미등록어인 경우의 음소들은 반드시 반음소 모델만으로 구성된다고 볼 수 없기 때문에 평가 기준이 될 수 없다. 따라서 등록어인 경우의 음소 단위 신뢰도 값을 기준으로 평가해야 한다. 이 표에서 알 수 있듯이, 평균적으로 /U/, /e/ 음소에서 반음소 모델과의 구분이 잘 되지 않음을 확인할 수 있다. 따라서 두 음소에 대한 모델과 반음소 모델은 더 개선할 필요가 있다. 한편, 표 5에서는 각 음소가 가지는 신뢰도 값들을 산술 평균낸 값이므로 표준편차가 어느 정도 있으므로 이를 줄이기 위한 노력도 해야 된다.

위에서 보여준 오류를 보완하기 위해서는 단위 단위의 신뢰도 및 음절 단위의 신뢰도 결합 사용, 음소 모델 및 반음소 모델의 형태 및 구조 개선 등이 요구된다.

표 4. 본 논문에서 제안한 발화 검증 방법의 성능
(반음소 모델 구현시 best Gaussian, 2nd best Gaussian, 3rd best Gaussian, 4th best Gaussian까지 포함)

Table 4. Performance of proposed utterance verification method (For anti-phoneme model including best Gaussian, 2nd best Gaussian, 3rd best Gaussian, 4th best Gaussian).

데이터 종류 / 임계값	시험용 데이터					평가용 데이터				
	CA	CR	FAI	FAO	FR	CA	CR	FAI	FAO	FR
-0.086	83.15	89.39	1.66	10.61	15.19	84.69	94.19	0.99	5.81	14.32
-0.090	84.63	88.26	1.85	11.74	13.52	86.66	91.92	0.99	8.08	12.35
-0.094	86.67	87.12	1.85	12.88	11.48	88.88	89.65	1.49	10.35	9.63
-0.098	87.97	84.28	2.40	2.40	9.63	89.87	87.12	1.49	12.88	8.64
-0.102	90.19	82.58	2.40	2.40	7.41	90.61	83.59	1.49	16.41	7.90

표 5. 각 음소별 음소 단위의 신뢰도 값

Table 5. Phoneme-based confidence score for each phoneme.

음소	음소단위 신뢰도 값	
	등록어	미등록어
a	-0.011	-0.175
ja	-0.25	-0.118
v	-0.049	-0.147
jv	-0.053	-0.182
o	-0.048	-0.140
jo	-0.068	-0.178
u	-0.088	-0.113
ju	-0.082	-0.118
U	*-0.099*	-0.131
i	-0.035	-0.149
we	-0.091	-0.149
wa	-0.062	-0.106
Wi	-0.022	-0.188
E	-0.071	-0.191
e	*-0.118*	-0.141
wE	-0.045	*-0.070*
wi	-0.044	-0.157
je	*	*
wv	-0.040	-0.157
g	-0.063	-0.133
G	-0.046	-0.164
n	-0.045	-0.173
d	-0.056	-0.157
D	-0.055	*-0.012*
r	-0.064	-0.144
m	-0.034	-0.150
b	-0.068	-0.129
B	-0.063	-0.145
s	-0.053	-0.142
S	-0.033	*-0.031*
N	-0.043	-0.124
z	-0.051	-0.182
Z	-0.070	-0.107
c	-0.069	-0.203
p	-0.061	-0.152
k	-0.052	-0.312
t	-0.065	-0.115
h	-0.046	-0.195

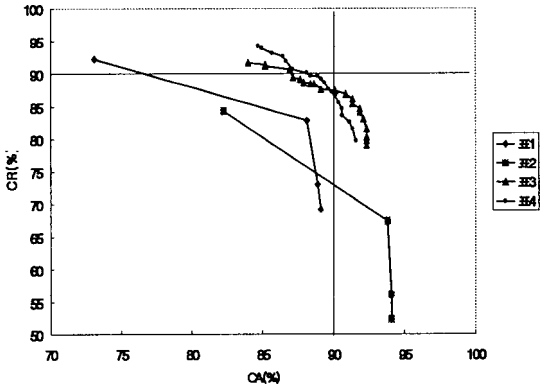

그림 6. CA-CR 곡선

Fig. 6. CA-CR curve.

V. 결 론

본 논문에서는 가변어휘 단어 인식에서의 미등록어 거절 성능을 향상시키는 효과적인 방법, 즉 최소 검증 오류를 보이는 반음소 모델과 단어 단위의 신뢰도를 제안하였다. 구체적으로 특별한 훈련 과정이 없이도 유사 음소 집합을 많이 포함시킨 반음소 모델을 제안하였다. 또한, 수식 (5)를 이용한 음소 단위의 null hypothesis와 alternate hypothesis의 비를 이용한 음소 단위의 신뢰도는 프레임 정규화보다 강인한 발화 검증 성능을 보여 주었으며, 음소 단위의 신뢰도를 단어 단위의 신뢰도로 변환한 수식 (4)는 등록어와 미등록어 사이의 분별력을 잘 표현해 주었다.

이와 같은 반음소 모델을 이용한 발화 검증 방법을 사용했을 때 CA는 약 89%, CR은 약 90%로써, 필러 모델을 사용한 핵심어 검출 방식을 이용해서 미등록어 거절시키는 기존 방법보다 성능이 월등히 우수함을 알 수 있었다.

향후 과제로써, 2단계 구조가 아닌 가변 어휘 단어 인식기에서 비터비 탐색시에 미등록어를 동시에 검출하는 1단계 구조를 연구할 계획이다.

감사의 글

본 연구는 1999년도 정보통신부 정보통신우수대학원 지원사업에 의거 1999년도 인하대학교 교내 연구비로 수행하였습니다.

참 고 문 헌

1. 김우성, 구명완, "반음소 모델링을 이용한 거절기능에 대한 연구", 한국음향학회지, vol. 18, no. 3, pp. 3-9, 1999.
2. Sunil K. Gupta and Frank K. Soong, "Improved Utterance Rejection Using Length Dependent Thresholds", ICSLP, 1998.
3. Li Jiang and Xuedong Huang, "Vocabulary-Independent Word Confidence Measure Using Subword Features", ICSLP, 1998.
4. Mazin G Rahim, Chin-Hui Lee, Biing-Hwang Juang and Wu Chou, "Discriminative Utterance Verification Using Minimum String Verification Error (MSVE) Training", ICASSP, 1996.
5. 김회린, 이항섭, "음성학적 지식 기반 변이음 모델을 이용한 가변 어휘 단어 인식기", 한국음향학회지, vol. 16, no. 2, pp. 31-35, 1997.
6. Hoi-Rin Kim, SiongHun Yi and Hang-Seop Lee, "Out-of-Vocabulary Rejection using Phone Filler Model in Variable Vocabulary Word Recognition", ICSP, vol. 1, pp. 337-339, 1999.

▲ 김 기 태 (Ki-Tae Kim)

1998년 2월 : 인하대학교 전자공학과 공학사
2000년 2월 : 인하대학교 전자공학과 공학석사
2000년 3월 ~ 현재 : LG전자 연구원
※ 주관심분야 : 음성인식, 오디오 코딩, 음성 코딩

▲ 문 광 식 (Kwang-Sik Moon)

1998년 2월 : 인하대학교 전자공학과 공학사
2000년2월 : 인하대학교 전자공학과 공학석사
2000년 3월 ~ 현재 : LG전자 연구원
※ 주관심분야 : 음성인식, 음성 코딩, 화자인식

▲ 정 재 호 (Jae-Ho Chung)
1982년 : University of Maryland (BSEE)
1984년 : University of Maryland (MSEE)
1990년 : Georgia Institute of Technology (Ph.D.)
1984년 ~ 1985년 : 미국 국방성 산하 해군 연구소, 신호처리실, 연구원
1991년 ~ 1992년 : AT&T Bell Laboratories, 음성신호처리 연구실, 연구원 (MTS)
1992년 ~ 현재 : 인하대학교 공과대학 전자공학과, (현)부교수

▲ 김 회 린 (Hoi-Rin Kim)
한국 음향학회지 제18권 제8호 참조

▲ 이 영 직 (Young-Jik Lee)
한국 음향학회지 제18권 제5호 참조