고립단어 인식 시스템에서의 거절기능 구현

An Implementation of Rejection Capabilities in the Isolated Word Recognition System

김 동 화*, 김 형 순**, 김 영 호***
(Dong-Hwa Kim*, Hyung-Soon Kim**, Young-Ho Kim***)

※본 연구에서 사용한 2가지 음성 데이터베이스는 한국전자통신연구원에서 제공한 것입니다.

요 약

고립단어 음성인식 시스템이 실용적이 되려면 인식 대상 이외의 단어를 거절할 수 있는 기능이 요구된다. 본 논문에서는 집단화된 음소 모델과 likelihood ratio에 의한 후처리 방법을 사용하여 거절기능을 구현하는 방법을 제안하였다. 기본적인 음성인식 시스템은 단어 단위 연속 HMM을 사용하였고, 6개의 집단화된 음소 모델들은 음성학적으로 균형잡힌 음성 데이터베이스를 이용하여 훈련된 45개의 문맥독립 음소 모델들로부터 통계적 방법에 의하여 생성되었다. 22개의 부서명칭을 대상으로 한 화자독립 고립단어 인식시스템에서 거절성능을 시험하여 본 결과, 가장 높은 확률값과 두 번째 높은 확률값을 가지는 후보단어들 간의 차이값에 의하여 거절기능을 수행하는 기존의 후처리 방법보다 성능이 향상됨을 알 수 있었다. 또한 이 집단화된 음소모델은 인식 대상 어휘가 다른 고립단어 인식 시스템에도 재훈련 없이 그대로 사용될 수 있다.

ABSTRACT

For the practical isolated word recognition system, the ability to reject the out-of-vocabulary(OOV) is required. In this paper, we present a rejection method which uses the clustered phoneme modeling combined with postprocessing by likelihood ratio scoring. Our baseline speech recognition system was based on the whole-word continuous HMM. And 6 clustered phoneme models were generated using statistical method from the 45 context independent phoneme models, which were trained using the phonetically balanced speech database. The test of the rejection performance for speaker independent isolated words recognition task on the 22 section names shows that our method is superior to the conventional post-processing method, performing the rejection according to the likelihood difference between the first and second candidates. Furthermore, this clustered phoneme models do not require retraining for the other isolated word recognition system with different vocabulary sets.

I. 서 론

음성은 사용의 편리함과 효율성으로 인간과 기계간의 중요한 의사소통의 수단으로 자리를 잡아가고 있다. 음성인식을 위하여 여러 가지 방법들이 사용되어지고 있는데 그 중에서 확률 모델인 HMM(Hidden Markov Model)[1, 2]이 가장 성공적으로 사용되고 있다. 음성인식 시스템 가운데 소규모 고립단어 인식 시스템이 그것의 효용성으로 인하여 많이 상용화되고 있다. 이러한 고립단어 인식 시스템의 경우 사용자들이 인식 대상 단어 이외의 말을 하는 경우가 자주 발생되는데 이런 경우를 대비하여 음성인식 시스템은 거절기능을 반드시 구비하여야 한다[3].

거절기능을 구현하는 방법은 거절을 위한 별도의 모델을 사용하는 방법과 적절한 후처리 과정을 통하여 인식결과를 확인하는 방법이 있다. 별도의 모델을 사용하는 방법에는 keyword spotting[4] 시스템에서 많이 사용되는 filler 모델링[5]과 반단어(anti-keyword)를 모델링하는 방법[6] 등이 있다. Filler 모델을 사용하는 경우에는 인식대상이 아닌 단어나 묵음을 어느정도 표현할 수 있는 filler

*밀양산업대학교 정보통신공학과 조교수
**부산대학교 전자공학과 조교수
***부산대학교 전자계산학과 부교수

접수일자 : 1997년 7월 28일

모델을 구성하며 이 모델을 어떻게 구성하는가에 따라 인식성능이 크게 좌우된다. 반단어의 모델링은 각단어를 모델링할 때 그 단어에 대한 반대단어를 함께 모델링을 하는 방법이다. 후처리 방법들에는 likelihood ratio에 의한 방법[7], 변별적 훈련과정을 사용하는 방법[5], 신경망을 이용하는 방법[8] 등이 있다.

본 논문에서는 HMM에 기반한 고립단어 인식 시스템에서 filler 모델링과 유사한 집단화된 음소 모델링과 likelihood ratio에 의한 후처리 방법을 함께 사용하여 거절기능을 구현하는 방법에 대하여 기술한다. 그리고 본 연구에서 제안한 방법과 기존의 후처리 방법 즉, Viterbi 디코딩 후의 확률값이 가장 높은 것과 두 번째로 높게 나타난 것의 차이값에 의하여 거절기능을 수행하는 방법을 대상으로 거절성능을 비교하였다.

Ⅱ. 고립단어 인식 시스템

음성인식을 위하여 HMM을 사용하는 방법에는 이산 HMM, 연속 HMM, 준연속 HMM 등이 있는데 본 연구에서는 인식성능이 가장 우수한 것으로 알려져 있는 연속 HMM을 사용하였으며, 실험을 위하여 상용화된 음성인식 개발도구인 HTK 2.0[9]을 사용하였다. 거절기능 구현 실험을 위한 기본적인 고립단어 음성인식 시스템으로는 단어 단위 HMM을 사용하였다. 이 장에서는 단어 단위 HMM을 사용한 고립단어 음성인식 시스템에 대하여 기술한다.

본 연구에서는 고립단어를 모델링하기 위하여 각 단어의 앞과 뒤에 묵음을 첨가하였으며 각각의 고립단어는 skip이 없는 left-to-right 방식의 연속 HMM으로 모델링하였다. 여기서 사용된 HMM은 18개의 상태 수와 2개의 mixture를 가지며, 이것은 여러 가지 상태 수와 mixture에 대하여 실험하여 본 후 인식률이 가장 높게 나오는 상태 수와 mixture 수를 선택한 것이다. 고립단어 인식실험은 한국전자통신연구원에서 제공한 부서명 음성 데이터베이스를 대상으로 하였으며, 다음 장에 기술될 음소 단위 HMM의 훈련용으로도 역시 동 연구원의 445 단어 음성 데이터베이스[10]를 사용하였다. 단어 단위 HMM을 사용한 고립단어 인식시스템의 훈련 및 인식 과정은 화자독립 방식으로 수행하였으며 HMM 초기화, Baum-Welch 재추정 및 비터비 디코딩 등으로 구성된다. HMM 초기화와 Baum-Welch 재추정에서 공분산 행렬(covarence matrix)에서의 대각원소인 분산과 mixture weight에 대하여 flooring을 수행하였으며 레이블링 정보로는 음성 데이터베이스와 함께 제공된 프레임 단위의 시작점/끝점 정보를 100ns의 단위로 변환하여 사용하였다. 재추정 사이클의 최대 횟수는 20회로 하였다. Viterbi 디코딩에서는 이미 훈련된 단어 HMM과 부서명으로 구성된 네트워크 화일 및 HMM 목록 화일을 사용하여 테스트 음성에 대하여 인식을 수행하고 그 결과를 대본(transcription) 파일에 출력한다. 이렇게 생성된 대본 파일과 레이블링 정보를 사용하여 최종적으로 인식률을 계산한다. 이러한 과정으로 부서명 데이터베이스에 대하여 남성화자 35명의 발성을 훈련용으로 사용하고 나머지 15명의 발성을 인식용으로 사용한 결과 99.14%의 인식률을 얻을 수 있었다.

Ⅲ. 거절기능 구현

거절기능을 구현하기 위하여 문맥독립 음소 모델들로부터 6개의 집단화된 음소 모델을 생성하고 이것을 일종의 filler 모델로써 사용하게 된다. 이 장에서는 문맥독립 음소모델의 구성 및 집단화 방법 그리고 이들을 사용하여 거절기능을 구현하는 방법에 대하여 기술한다.

3.1 음소 모델의 구성 및 집단화 방법

음소 단위 HMM의 구조는 3개의 상태 수와 1개의 mixture를 가지며 skip이 없는 연속 HMM이다. 음소 모델들을 구성하기 위하여 인식 대상 어휘와는 무관한 음성학적으로 균형잡힌 445단어를 사용하였으며 22명이 2회씩 발성한 것 중에서 1회 발성분만 훈련에 사용하였다. 재정렬과 재추정 과정을 반복하여 묵음 모델을 포함한 46개의 문맥독립 음소 모델들을 생성하였다.

후처리를 통한 거절기능을 구현하기 위하여 45개의 음소 모델들(묵음 모델은 제외)로부터 통계적 방법에 의한 monophone clustering[11] 알고리즘을 사용하여 6개의 집단화된 음소 모델을 생성하였다. Monophone clustering을 위한 거리척도는 다음 식 (1)과 같다.

$$D(P_i, P_j) = \sum_{d=1}^{N} D_d(P_i, P_j) \tag{1}$$

여기서 P_i, P_j는 각각 i, j번째 음소를 나타내고, N은 음소 모델의 상태 수를 나타내며 $D_d(P_i, P_j)$는 두 음소의 각 상태간의 거리로서 다음 식 (2)와 같이 주어진다.

$$D_d(P_i, P_j) = \frac{1}{V} \sum_{k=1}^{V} \frac{(\mu_{idk} - \mu_{jdk})^2}{\sigma_{idk} \cdot \sigma_{jdk}} \tag{2}$$

여기서 V는 관찰 벡터의 차수를 나타내고 μ_{idk}와 σ_{idk}는 각각 i 번째 음소의 d 번째 상태의 평균과 표준편차를 나타낸다. 이상의 거리척도를 사용하는 K-means 알고리즘을 이용하여 집단화를 수행하고 이 집단화 정보를 이용하여 6개의 새로운 음소 모델을 재훈련 과정을 통하여 다시 생성하였다.

3.2 거절기능 구현

지금까지 일반적으로 사용되어온 후처리 방법에 의한 거절기능 구현방식은 Viterbi 디코딩의 결과 첫 번째 후보와 두 번째 후보의 likelihood 차이를 계산하여 특정 문턱값 미만이 되면 즉, 식 (3)을 만족하면 거절하는 것이다. 식(3)에서 C1과 C2는 첫 번째와 두 번째 후보에 해당

최적 단어 상의 HMM을 의미하며 θ_1은 거절을 위한 문턱값이다.

$$\log P(O|C_1) - \log P(O|C_2) < \theta_1 \quad (3)$$

본 연구에서 제안하는 집단화된 음소 모델을 이용한 후처리 방법에서는 식 (4)를 만족하면 거절기능을 수행하게 된다. 여기서 $C1$은 단어 단위 HMM에서의 첫 번째 후보이고, $CP1$은 인식 대상 단어에 대한 6개의 집단화된 음소 모델들로 구성된 most likely sequence를 나타내며 θ_2는 역시 거절을 위한 문턱값이다.

$$|\log(O|C_1) - \log(O|CP1)| < \theta_2 \quad (4)$$

일반적으로 문턱값을 크게 하면 인식대상 어휘에 대한 잘못된 거절률(false rejection rate)이 높아지며, 문턱값을 작게 하면 반대로 인식대상이 아닌 어휘를 받아들이는 비율(false acceptance rate)이 높아진다. 여기서 사용된 6개의 clustered 음소 모델은 인식 대상 어휘와는 독립적으로 생성되었기 때문에 인식 대상 어휘가 바뀌더라도 별도의 훈련 과정 없이 그대로 사용될 수 있다는 장점을 가진다.

Ⅳ. 실험 및 결과

4.1 실험 환경 및 데이터

본 실험은 sun SPARC20 시스템에서 HTK 2.0을 사용하여 수행되었다. 거절기능 실험을 위하여 별도의 거절용 음성데이터를 준비하지 않고 편의상 22개의 ETRI 부서명을 다음 표 1과 같이 임의적으로 인식 대상 어휘와 인식 대상이 아닌 어휘로 구분하였다.

표 1. 어휘목록
Table 1. Vocabulary list

인식대상 어휘(11개)	비인식대상 어휘(11개)
총무부	건설과
운영관리실	설비과
총무과	영선과
자산관리과	근로복지실
안전관리과	근로과
예비군대대본부	복지과
재무관리실	서울사무소
회계과	인력개발부
내자과	인력계획과
외자과	인사과
건설관리실	연수실

ETRI 445 단어 및 부서명 데이터베이스는 모두 16 kHz sampling rate, 16 bit 양자화, 7 kHz LPF를 거쳐 구축된 것이며, 이 음성 데이터를 사용한 본 실험에서의 훈련 및 인식을 위한 전처리 과정으로서 preemphasis 계수는 0.97로 하였으며 해밍창을 사용하여 20ms 구간과 10ms 주기로 분석하였다. 음성특징 파라미터는 12차 캡스트럼, 12차 델타 캡스트럼, 로그 에너지와 델타 로그 에너지를 사용하였다.

4.2 실험 결과

거절기능을 가진 음성인식 시스템의 성능을 평가하기가 쉽지 않다. 왜냐하면 거절기능을 수행하지 않고 인식대상 어휘만에 대한 인식률과 인식대상이 아닌 어휘에 대한 높은 거절률이 요구되는 경우의 인식대상 어휘에 대한 거절률 사이에는 trade-off 관계가 있기 때문이다[12]. 본 연구에서는 화자독립 고립단어 인식시스템에서 거절성능을 평가하기 위하여 다음과 같은 2가지 항목들에 대하여 분석하였다.

① 인식대상 어휘에 대한 거절률(false rejection rate)에 따른 인식대상이 아닌 어휘에 대한 거절률(그림 1)
② 인식대상 어휘에 대한 거절률에 따른 거절되지 않은 인식대상 어휘의 인식률(그림 2)

그림 1에서 top12는 첫 번째 후보단어와 두 번째 후보단어의 비터비 출력값의 차이에 의하여 거절을 수행하는 기존의 방법이고 clust는 본 연구에서 제안한 방법이다. 모든 구간에서 제안된 방법의 성능이 우수함을 볼 수 있다. 인식대상 어휘에 대한 인식률은 거절기능을 수행하기 전에는 98.8%이었으며, 그림 2에서 처럼 거절기능을 수

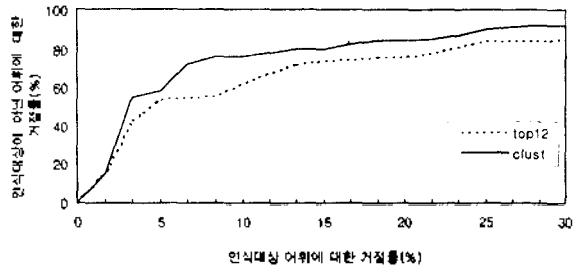

그림 1. 인식대상 어휘의 거절률에 대한 인식대상이 아닌 어휘의 거절률
Fig. 1 Keyword rejection versus non-keyword rejection rate

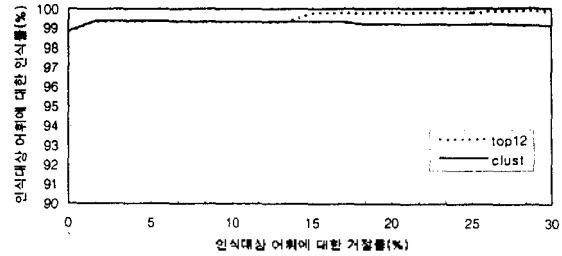

그림 2. 인식대상 어휘의 거절률에 따른 인식률
Fig. 2 Keyword rejection rate versus accuracy

핵함에 따라 거절되지 않은 인식대상 어휘의 인식률이 기존의 방법은 98.8%에서 99.9%이고 제안된 방법은 98.8%에서 99.3%사이의 분포를 나타내었다. 두 방법 모두에서 교체오류(substitution error) 없이, 1-2개로서 [illegible] 인 경우에 인식률을 비교하는 것은 현실적으로 별 의미가 없다.

V. 결 론

본 연구에서는 실용적인 화자독립 고립단어 인식 시스템의 구현을 위하여 집단화된 음소 모델과 likelihood ratio에 의한 후처리 방법을 함께 사용하여 거절기능을 수행하는 방법을 제안하였다. 기본적인 음성인식 시스템은 단어 단위 연속 HMM을 사용하였고, 집단화된 음소 모델들은 음성학적으로 균형잡힌 445 단어 음성 데이터베이스를 사용하여 훈련된 45개의 음소 모델들로부터 통계적 방법에 의한 집단화와 재훈련 과정을 통하여 생성되었다. 22개의 부서명칭을 인식대상 어휘와 비인식대상 어휘로 구분하여 거절성능을 시험하여 본 결과, 기존의 likelihood ratio에 의한 후처리 방법보다 제안된 방법이 우수함을 알 수 있었다. 또한 집단화된 음소 모델은 인식 대상 어휘와는 독립적으로 생성되었기 때문에 인식 대상 어휘가 다른 여러 가지 소규모 고립단어 인식 시스템에도 재훈련이 필요없이 그대로 사용될 수 있다.

앞으로 제안된 방법과 신경망을 결합하여 다시 거절성능을 시험해 볼 예정이며, 아울러 이 방법을 잡음환경 음성인식 시스템에도 사용해 보고자 한다.

참 고 문 헌

1. L. R. Rabiner, B. H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.
2. X. D. Huang, Y. Ariki, M. A. Jack, *Hidden Markov Models for Speech Recognition*, Edinburgh, 1990.
3. 구명완, "신경망을 이용한 음성인식 거절기능 구현," 제13회 음성통신 및 신호처리 워크샵 13권 1호, pp. 207-211, Aug., 1996.
4. 김형순, "Keyword Spotting 기술," 한국통신학회지 제 11권 9호, pp. 57-65, Sep., 1994.
5. R. A. Sukkar, et al., "A two pass classifier for utterence rejection in keyword spotting," *Proc. ICASSP-93*, pp. 451-454, 1993.
6. M. G. Rahim, et al., "Robust utterence verification for connected digits recognition," *Proc. ICASSP-95*, pp. 285-288, 1995.
7. R. C. Rose, et al., "A hidden Markov model based keyword recognition system," *Proc. ICASSP-90*, pp. 129-132, 1990.
8. D. P. Morgan, et al, "A keyword spotter which incorporates neural networks for secondary processing," *Proc. of ICASSP-90*, pp. 113-116, 1990.
9. Steve Young, *The HTK Book*, Entropic, 1996.
10. [illegible]
11. [illegible]
12. Sari Accaino, et al., "Rejection Capabilities For HMM-based Speech Recognizers," *Proc. EUROSPEECH'95*, pp. 2115-2118, 1995.

▲김 동 화(Dong-Hwa Kim)

1982년 2월 : 부산대학교 수학교육과 졸업(이학사)
1989년 2월 : 부산대학교 전자계산학과 졸업(이학석사)
1995년 2월 : 부산대학교 전자계산학과 박사과정 수료
1993년 3월~현재 : 밀양산업대학교 정보통신공학과 조교수

※주관심분야 : 음성인식, 휴먼 인터페이스

▲김 형 순(Hyung-Soon Kim)

현재 : 부산대학교 전자공학과 조교수
한국음향학회지 제16권 5호 참조

▲김 영 호(Young-Ho Kim)

1982년 2월 : 서울대학교 컴퓨터공학과 졸업(공학사)
1984년 2월 : 서울대학교 컴퓨터공학과 졸업(공학석사)
1991년 2월 : 서울대학교 컴퓨터공학과 졸업(공학박사)
1995년 1월~1996년 1월 : 미국 University of Pennsylvania 방문연구
1989년 3월~현재 : 부산대학교 전자계산학과 부교수