한국음향학회지 제25권 제2호 pp. 96-99 (2006)

정규화신뢰도 기반 가변어휘 고립단어 인식기의 거절기능 성능 분석

Rejection Performance Analysis in Vocabulary Independent Speech Recognition Based on Normalized Confidence Measure

최 승 호*
(Seung Ho Choi*)

*동신대학교 멀티미디어학과
(접수일자: 2006년 1월 11일; 채택일자: 2006년 2월 15일)

고립단어 인식기의 오 인식 단어를 거절하기 위한 방법으로 정규화 신뢰도가 제안되어 논문 [1-2]에서 성공적으로 적용된 바 있다. 그러나 정규화 신뢰도의 성능 측정을 위해 고정된 단어 셋을 대상으로 실험을 하였다.
본 논문에서는 정규화 신뢰도를 가변어휘 음성인식 영역에 적용하여 신뢰도의 거절성능을 밝히고 특히, 벡터양자화기를 이용하여 미 출현 트라이 폰의 문제를 극복하는 방법을 제안한다. 이때 정규화 신뢰도는 트라이 폰 신뢰도들의 통계적 특징(평균과 표준편차)을 사용한다.
가변어휘 인식실험 결과 음소 단위의 정규화 방법이 트라이 폰 기반 정규화 방법에 비하여 우수한 성능을 보였으며 이러한 결과는 논문 [1-2]의 결과와는 상이한 것으로 트라이 폰 기반 정규화 방법이 미 출현 트라이 폰에 대하여 강인하지 못하다는 점을 시사하고 있다.
따라서 정규화 신뢰도가 음소 또는 트라이 폰에 상관없이 기존 신뢰도인 RLJC 신뢰도 [3]에 비하여 우수한 성능을 보였으며 가변어휘 인식에서도 동작함을 확인 할 수 있었다.

핵심용어: 가변어휘 음성인식, 오 인식 단어 거절, 정규화 신뢰도

투고분야: 음성처리 분야 (2.5)

Kim et al. proposed Normalized Confidence Measure (NCM) [1-2] and it was successfully used for rejecting mis-recognized words in isolated word recognition. However, their experiments were performed on the fixed word speech recognition.
In this paper we apply NCM to the domain of vocabulary independent speech recognition (VISP) and shows the rejection performance of NCM in VISP. Specially we propose vector quantization (VQ) based method for overcoming the problem of unseen triphones. It is because NCM uses the statistics of triphone confidence in the case of triphone-based normalization.
According to speech recognition experiments phone-based normalization method shows better results than RLJC [3] and also triphone-based normalization approach. This results are different with those of Kim et al [1-2]. Concludingly the phone-based normalization shows robust performance in VISP domain.

Keywords: Vocabulary Independent Speech Recognition, Rejection of Mis-recognized Words, Normalized Confidence Measure

ASK subject classification: Speech Signal Processing (2.5)

책임저자: 최 승 호 (shchoi@dsu.ac.kr)
520-714 전남 나주시 대호동 252번지
동신대학교 멀티미디어학과
(전화:061-330-3194; 팩스:061-330-2909)

I. 서 론

음성은 인간이 사용하는 가장 편리하고 효과적인 정보 교환 수단이다. 따라서 지난 30여 년간 인간과 기계와의 인터페이스 수단으로써 음성인식의 연구가 지속적으로 이루어져왔다. 그러나 음성인식이 우리의 일상생활까지 일반화되지 못하고 단지, 몇몇 고립단어 인식기가 성공적으로 상용화되어 사용되고 있을 뿐이다.

음성 인식기가 실제 서비스에서 사용되지 못하는 이유는 잡음과 채널왜곡 등에 의한 낮은 음성인식 성능, 오인식 단어에 대한 부적절한 처리 등이 있다. 특히, 오 인식 단어 거절 성능은 사용자의 불편함을 해소하기 위한 중요 요소이다.

신뢰도는 거절기능의 척도로 사용되며 일반적으로 음소단위 신뢰도를 구한 후 이를 가중평균을 취하여 단어 신뢰도를 구한다[3]. 최근, 이러한 신뢰도 계산 방법의 향상을 위해 신뢰도 정규화가 제안되었는데 이를 정규화 신뢰도 (normalized confidence measure)라 한다[1-2]. 정규화 신뢰도는 음소 또는 트라이 폰 단위 신뢰도의 정규화를 얻기 위해 각 단위 신뢰도들의 통계적 성질인 평균과 표준편차를 구하고 각 단위 신뢰도가 동일한 분포를 갖도록 정규화 한다. 정규화 신뢰도의 검증을 위해 고립단어 음성 인식기에 적용하여 발표되었다[1-2].

본 논문에서는 정규화 신뢰도를 가변어휘 음성 인식기에 적용하고 인식실험을 하였다. 특히, 가변어휘 음성 인식기에서는 인식기 학습 시 나타나지 않는 미 출현 트라이 폰 (unseen triphone: UT)이 존재하기 때문에 트라이 폰 기반의 정규화를 수행할 수 없다. 그 이유는 인식기가 UT에 대한 신뢰도의 분포를 가지고 있지 않기 때문이다.

따라서 본 논문에서는 이 문제의 해결 방법으로 첫째, 벡터양자화 기반 집단화를 이용하여 UT의 신뢰도 분포를 추정하는 방법을 제안하고 둘째, 중규모급의 가변어휘 인식기를 사용하여 정규화 신뢰도의 성능을 검증한다.

II. 정규화 신뢰도와 미 출현 트라이 폰 처리 방법

2.1. 정규화 신뢰도

신뢰도를 비교 분석하기 위해 사용된 신뢰도는 Rahim 등에 의하여 제안된 RLJC 신뢰도[3]이다. 이 방법은 음소단위 신뢰도를 구하고 이 신뢰도에 가중평균을 취하여 단어단위 신뢰도를 계산한다.

음소단위 신뢰도는 음성 인식기의 출력인 음소 확률과 인식된 음소의 반 음소 (anti-phone) 확률과의 비로 정의되며 식 (1), (2), (3)와 같다.

$$\log pr_a = \frac{1}{M}\sum_{i=0}^{M-1}\log pr_a, \tag{1}$$

$$cm_p = \frac{\log pr_p - \log pr_a}{|\log pr_p|} \tag{2}$$

$$CM = \frac{1}{f_{cm}}\log\left(\left(\sum_{p=0}^{n_p-1}\exp(f_{cm}\cdot cm_p)\right)/n_p\right) \tag{3}$$

여기에서 $\log pr_a$는 반 음소 모델의 평균 로그 확률, $\log pr_p$는 단어구성 음소 모델의 로그 확률, n_p는 단어구성 음소 수, M은 반 음소 모델의 수, cm_p는 음소단위의 신뢰도, f_{cm}는 음의 값을 갖는 가중치이다.

일반적인 반 음소는 전체 음소 집합에서 인식된 음소를 제외한 나머지 음소를 사용하게 되며 반 음소의 확률 계산을 위해서는 음소HMM (hidden Markov model)을 이용한다.

RLJC 신뢰도는 음소마다 신뢰도의 분포가 동일하지 않기 때문에 단어마다 신뢰도의 분포가 다르고 단어마다 거절 성능이 균일하지 않는 문제점을 갖고 있다. 이것을 해결하기 위한 방법이 정규화 신뢰도이다[1-2].

참고 논문[1-2]에서 제안된 신뢰도는 음소 또는 트라이 폰 단위로 신뢰도의 분포가 동일해지도록 정규화 하였다. 이러한 정규화 신뢰도를 고립단어 인식기에 적용함으로써 거절성능을 검증하였다. 이때 정규화 신뢰도는 식 (4), 식 (5)과 같다.

$$ncm_p = \frac{cm_p - \mu_{nPhone,p}}{\sigma_{nPhone,p}} + \alpha \tag{4}$$

$$NCM = \frac{1}{f_{ncm}}\log\left(\left(\sum_{p=0}^{n_p-1}\exp(f_{ncm}\cdot ncm_p)\right)/n_p\right) \tag{5}$$

여기에서 f_{ncm}은 음소단위 신뢰도의 가중치, cm_p는 음소단위 신뢰도, ncm_p는 정규화 음소단위 로그확률, α는 음소단위 신뢰도의 정규화 가중치, $\mu_{nPhone,p}$는 음소단위 신뢰도의 평균, $\sigma_{nPhone,p}$는 음소단위 신뢰도의 표준편차, $nPhone$은 n값에 따른 음소 ($n=1$), 또는 트라이 폰 ($n=3$)을 나타낸다.

이와 같이 정규화 신뢰도는 음소 신뢰도를 음소단위마다 신뢰도의 평균과 표준편차로 구해진다.

2.2. 가변어휘 인식기에서 미출현 트라이 폰의 처리 방법

2.1절의 식 (4)과 (5)의 정규화 신뢰도는 고립단어 인식기 거절기능의 성능을 향상시킬 수 있지만 가변어휘 인식기에서는 음성인식기의 학습 시 사용했던 데이터베이스에 출현하지 않은 트라이 폰이 가변어휘 인식에서 나타날 수가 있기 때문에 UT를 해결할 수 있는 새로운 방법이 제안되어야 한다. 가변어휘 음성인식에서 UT는 상태묶음으로 HMM 모델을 만들면서 결정나무 (decision tree)를 이용한다[4].

따라서 정규화 신뢰도의 문제점은 HMM 모델 자체가 아니고 새로 정의된 UT의 신뢰도에 대한 통계적 분포의 예측이다.

본 논문에서는 이러한 UT 신뢰도의 평균과 표준편차 분포를 예측하기 위해 벡터양자화 기반의 집단화 방법을 제안하여 정규화 신뢰도의 문제점을 해결하고자 한다.

벡터양자화기의 학습 방법은 Linde 등이 제안한 LBG 알고리듬으로 k-means 집단화와 분할기법을 사용하여 최적의 집단 대표를 구하는 방법이다. 음성코딩과 패턴인식 등에서 국부적 최적화 문제를 해결하기 위한 방안으로 k-means 집단화 방법이 널리 사용되고 있다[5-6].

본 논문에서는 LBG 알고리듬으로 트라이 폰을 집단화하고 집단별로 신뢰도의 평균과 표준편차를 공유한다.

아래와 같은 1)~3)과정을 통하여 구해진 트라이 폰 집단과 대표 그리고 각 집단마다 계산된 신뢰도의 평균과 표준편차를 이용하면 가변어휘 음성인식 시 UT 신뢰도의 평균 및 표준 편차를 예측할 수 있다. 즉, UT에 대한 HMM 모델이 결정되면 벡터양자화로 소속 집단을 결정하고 공유된 신뢰도의 통계적 성질을 이용한다.

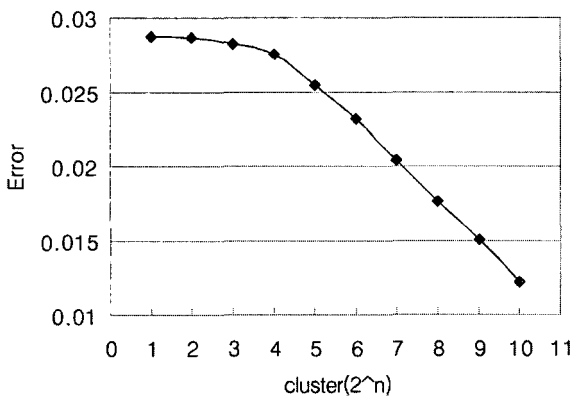

그림 1. 집단 수에 따른 평균에러
Fig. 1. Average of Cluster Number.

1) 각 트라이 폰을 1 mixture 3 상태의 HMM 학습하여 트라이 폰 마다 3개의 특징벡터를 구한다.
이때 i 번째 트라이 폰의 특징을 표현하는 벡터는 $V_i = v_{mean1}, v_{mean2}, v_{mean3}$ 이고 $v_{mean\,j}$ 는 j번째 상태의 가우시안 모델의 평균벡터이다.

2) V_i를 모아서 유클리디안 거리함수를 이용하여 LBG기반 집단화를 수행한다.

3) 트라이 폰의 각 집단에서 공유된 트라이 폰 단위 신뢰도의 평균과 표준편차를 각각 구해지며 이는 각 음성인식기를 이용한 레이블링 처리 과정에서 얻을 수 있다.

본 논문에서는 상태 당 26차의 벡터를 구했으며 Mel-cepstrum, 에너지, delta 파라미터 등의 음성특징파라미터를 사용하였다.

2.2.1. 집단화 알고리즘

그림 1은 집단 수에 따른 양자화 오차를 나타낸 것으로써 집단수가 감소되면 평균 에러는 증가하게 된다. 즉 1) 트라이 폰 개수를 줄이면 집단화되는 트라이 폰 수가 증가하면서 에러 값이 증가되고 2) 이와 반대로 집단화 수가 증가되면 평균 에러는 감소하게 된다.

본 논문에서는 트라이 폰 수를 2347개, 최대 집단 수 1024개로 제한하여 실험 하였다.

III. 실험 및 분석

3.1. 가변어휘 음성인식 시스템

3.1.1. 음성 인식기 학습

음성 인식기 학습에는 총 400 단어의 지명 음성 DB를 사용하였으며 각 단어는 72명의 화자로부터 수집하였고, A/D변환은 8khz 샘플링과 16bit 양자화 기반으로 수행되었다. 음성 수집은 조용한 연구실환경에서 일반 컨덴서 마이크를 이용하여 구축하였다.

트라이 폰 모델 학습은 200단어와 52명의 음성 DB를 대상으로 수행하였으며, 출현 트라이 폰 수는 2347개이다.

HMM학습은 HTK (hidden Markov model tool kit) 2.1을 사용하여 ergodic 3 상태 HMM의 상태 묶기 (state tying)를 이용하였다.

3.1.2. 인식기 디코더

실험에 사용된 가변어휘 음성 인식 디코더는 음운 변동기를 내장하고 있으면서 HTK의 상태 묶기 결정트리로 UT에 대한 모델을 구현하고 디코더는 캠브리지대에서 개발된 토큰 패싱 알고리듬을 이용하였다[7].

인식된 단어에 대한 후처리에서 신뢰도는 인식단어에 대해 반 음소 모델을 구성하고 반 음소 모델의 확률과의 상대적 유사도를 계산하여 발화의 수락여부를 판단한다.

3.1.3. 거절 성능 평가 기준

거절기능의 성능 평가를 위하여 *MDR* (miss-detection ratio)과 FAR (false alarm ratio)을 기준으로 정하였다. MDR은 테스트 단어에서 출현한 인식 단어를 인식기가 제대로 검출하지 못한 단어수를 전체 인식단어와 비교한 값이고 FAR은 테스트 단어에서 출현한 인식 단어를 인식기가 오 인식 한 단어수를 전체 인식단어와 비교한 값이다. 이때 FAR은 인식 단어 당 FA의 출현횟수로 정규화 하였다.

다음 식 (6)과 식 (7)은 실험에 사용된 MDR과 FAR를 식으로 나타낸 것이다.

$$MDR = \text{미검출단어수}/\text{전체인식단어수} \quad (6)$$

$$FAR = \text{오인식단어수}/\text{전체인식단어수} \quad (7)$$

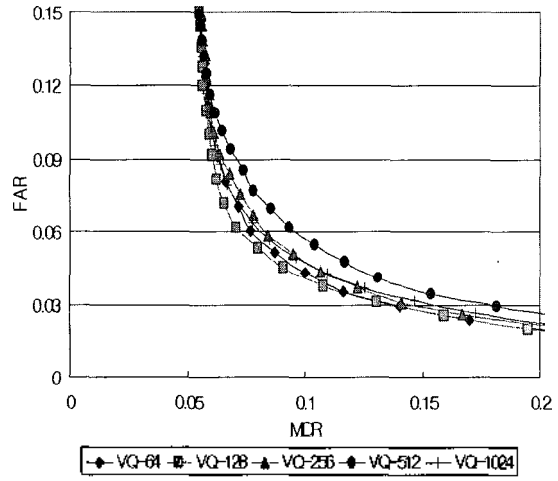

그림 2. 집단 수에 따른 거절 성능
Fig 2. Rejection Performance of Cluster Number.

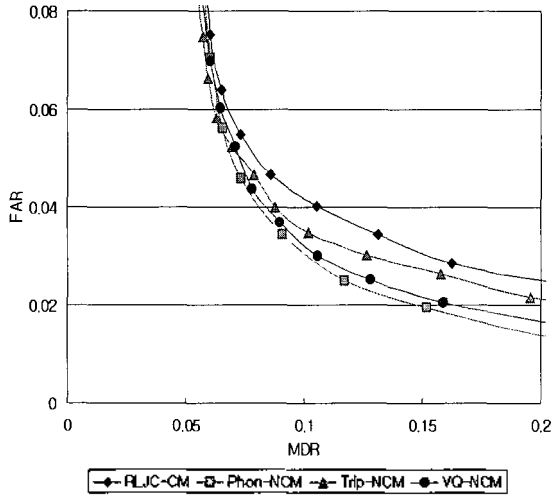

그림 3. 각 신뢰도의 거절 성능
Fig 3. Rejection Performance of Confidence Measures.

3.2. 거절 성능 실험 결과

가변어휘 음성 인식기의 거절 성능을 평가하기 위하여 음성인식 HMM 모델 학습에 참여하지 않은 200단어를 대상으로 인식 실험 및 거절 성능을 검증 하였으며 그 중 실험은 18명을 대상으로 하였다. 200단어 중 100단어는 인식 리스트에 부가하여 인식 대상 단어로 나머지 100단어는 사기단어 (imposter word)로 사용하였다.

거절 성능 평가는 먼저 벡터양자화기의 집단 수에 따라 수행하였으며 이 실험은 고립단어 인식기를 대상으로 이루어졌다.

집단 수에 따른 거절 성능의 결과는 그림 2에 나타내었으며 이를 분석해보면 집단의 수를 계속 증가시킬 때 거절 성능이 지속적으로 향상되지 않음을 확인 할 수 있다. 그리고 집단 수가 VQ-128일 때 최적의 성능을 보였으며, 집단 수가 VQ-256, 512, 1024로 커질 경우 성능이 저하됨을 알 수 있다.

가변어휘 인식의 거절 성능을 VQ-128기반의 RLJC-CM[3], Phon-NCM, Trip-NCM, VQ-NCM 등으로 구분하여 신뢰도의 성능을 그림 3에 나타내었다. Phon-NCM은 부 단어 단위 (subword unit)들의 신뢰도 정규화시 트라이 폰이 아닌 음소를 이용하였고, Trip-NCM은 집단화를 통한 신뢰도 통계 모델을 공유하지 않기 때문에 미 출현 트라이 폰이 발생한 경우 기존 트라이 폰 모델에 유클리디안 거리가 가장 가까운 트라이 폰 정규화 모델을 사용하였다. 그 실험 결과 가장 우수한 성능을 보인 것은 음소단위의 정규화 신뢰도인 Phon-NCM으로 나타났다. 이는 고정단어를 대상으로 한 인식실험 [1-2] 결과와 비교해 보면 다른 결과임을 알 수 있다. 참고문헌[1-2]에 의하면, 트라이 폰 단위의 신뢰도가 음소단위 정규화에 비하여 우수한 성능을 나타낸 것으로 보고되어 있다.

가변어휘 인식실험에서 트라이 폰 단위 정규화 방법의 거절성능이 음소단위 정규화에 비하여 낮은 이유는 트라

이 폰 신뢰도의 통계특성이 과도하게 학습데이터에 편중되게 추정되었기 때문이라고 판단된다.

VQ-NCM의 경우 Phon-NCM과 거의 유사한 성능을 보였지만 Phon-NCM 방법의 간편함을 고려할 때 가변어휘 음성인식에서는 음소단위 정규화 신뢰도인 Phon-NCM을 사용하는 것이 바람직하다고 판단된다.

그리고 신뢰도의 원 방법인 RLJC-CM [3]에 비하여 NCM의 거절성능이 향상됨을 그림 3에서 알 수 있다.

VI. 결 론

본 논문에서는 정규화 신뢰도 기반의 가변어휘 단어인식기의 거절가능 성능을 분석하였다. 가변어휘 인식에서는 미 출현 트라이 폰의 영향으로 트라이 폰 기반 정규화 신뢰도를 직접적으로 고려하기 힘들기 때문에 VQ를 이용한 신뢰도 통계 특성 공유방법을 제안한 결과 기존 고정단어 인식기의 경우와는 달리 음소단위의 정규화 신뢰도를 사용하는 것이 가장 강인함을 확인할 수 있었으며, VQ를 이용한 트라이 폰 정규화 신뢰도는 유사한 성능을 보였다.

따라서 본 논문에서는 정규화 신뢰도가 가변어휘 단어인식기에서 성공적으로 동작할 수 있음을 확인하였다.

감사의 글

본 논문은 2004년도 동신대학교 교내학술연구비 지원에 의하여 수행되었습니다.

참 고 문 헌

1. J. Kim, J Lee, S. Choi, "Hybrid Confidence Measure for Domain-Specific Keyword Spotting", Proc. of IEA/AIE, 15, 736-745, 2002
2. 김철, 이경록, 김진영, 최승호, 최승호, "정규화 신뢰도를 이용한 핵심어 검출 성능 향상", 한국음향학회지, 21 (4), 380-386, 2002.
3. M.G. Rahim, C.H. Lee, B.H. Juang, W. Chou, "Discriminative utterance verification using minimum string verification error (MSVE) training", ICASSP-96, 3585-3588, May, 1996
4. S. Young, HTK Book, ver 2.1, Cambridge University, 1997.
5. Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for Vector Quantizer design", IEEE Trans. on Communications, COM - 28 (1), 84-95, January, 1980.
6. E. Bocchieri, "Vector Quantization for the Efficient Computation of Continuous Density Likelihoods", in Proc. of the IEEE Int. Conf. Acoustic, Speech, Signal Processing, 692-695, April, 1993.
7. S.J. Young, N.H. Russell, J.H.S. Thornton, "Token Passing a simple conceptual model for connected speech recognition systems", Technical report of Cambridge University Engineering Department, TR38, July, 1989.

저자 약력

• 최 승 호 (Seung Ho Choi)
한국음향학회지 제21권 제4호 참조
1992년 3월~현재: 동신대학교 멀티미디어학과 교수