CM 알고리즘을 이용한 핵심어 검출 시스템의 인식률 향상에 관한 연구

원종문*, 이정숙, 김순협
광운대학교 컴퓨터공학과

A Study on the Recognition-Rate Improvement by the Keyword Spotting System using CM Algorithm

Jong-Moon Won*, Jung-Suk Lee, Soon-Hyob Kim
Kwangwoon University
E-mail : stblack@korea.com, netspee@explore.gwu.ac.kr, kimsh@daisy.gwu.ac.kr

요 약

본 논문은 중규모 단어급의 핵심어 검출 시스템에서 인식률 향상을 위해 미등록어 거절(Out-of-Vocabulary rejection) 기능을 제어하기 위한 연구이다. 이것은 핵심어 검출기에서 인식된 결과를 확인하는 과정으로 검증 시스템이 구현되기 위해서는 매 음소마다 검증 기능이 필요하고, 이를 위해서 반음소(anti-phoneme model) 모델을 사용하였다. 검증의 역할은 인식기에서 인식된 단어가 등록어인지 미등록어인지 판별하는 것이다. 단어 인식기는 비터비 탐색을 하므로, 기본적으로 단어단위로 인식을 하지만 그 인식된 단어는 내부적으로 음소단위로 인식된다. 따라서, 최소 검증 오류를 갖는 반음소 모델을 사용하고, 이를 이용하여 인식된 음소 단위들을 각각의 반음소 모델과 비교하여 통계적인 방법에 의해 신뢰도를 구한다. 이 음소단위의 신뢰도를 단어 단위의 신뢰도로 환산하기 위해서 음소단위를 평균 내는 방식을 취한다. 이렇게 함으로서, 등록어와 미등록어 사이의 분별력을 크게 하여 향상된 인식 성능을 얻었다.

1. 서론

음성은 특성상 인간의 가장 기본적인 정보교류 형태 중의 하나로서 음성을 이용한 인식시스템 개발은 인간과 기계와의 친화력 향상에 가장 중요한 요인이다.

Confidence Measure는 인식 시스템에서 등록이 안된 음성을 발성하면 이를 처리할 수 없다는 단점을 지니게 되므로 사용자는 정해진 등록어만을 사용해야 하는 제약을 받는다. 본 논문은 핵심어 모델과 필러 모델을 사용하는 연결단어 인식 알고리즘을 기반으로 한 핵심어 검출방식보다 성능이 우수한 거절기능 검증 방식을 제안하였으며, 거절기능의 역할은 연결단어 인식기에서 인식된 단어가 등록어인지 미등록어인지 판별하는 것이다. 음성인식의 많은 응용분야에서 통계적인 hypothesis 테스트를 이용하여 핵심어 검출과 거절기능이 실행된다. 일반적으로 유사도 비를 이용한 테스트를 많이 사용하는데, 그 방법은 입력단어가 등록단어라고 가정하는 null hypothesis와 미등록어라고 가정하는 alternative hypothesis 와의 비를 이용하는 것이다. alternative hypothesis 는 2가지의 카테고리를 포함하고 있다. 즉, 미등록어와 잘못 인식된 등록어가 이에 해당한다.

인식 시스템은 비터비 탐색을 하기에 기본적으로 단어단위로 인식되지만, 그 인식된 단어는 내부적으로 음소단위로 인식이 된다. 따라서, 최소 검증 오류를 갖으며 훈련이 필요 없는 반음소 모델을 제안하고, 이를 이용하여 인식된 음소단위들을 각각의 음소단위의 신뢰도를 환산하기 위해서 음소단위의 신뢰도를 평균내는 방

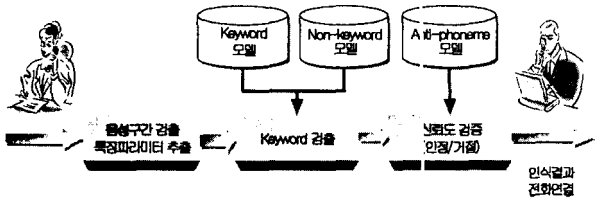
식을 취한다. 이렇게 함으로서 등록어와 미등록어의 분별력을 크게 하였다. 그리고 검출 성능의 향상을 위하여 Gaussian mixture model(GMM)을 이용한 반음소 모델(Anti-Phoneme model)을 구성한 다음 검출 속도의 향상을 위해 유사음소들을 clustering 하여 계산량을 감소시켜 인식 속도를 향상 시켰다.

2. 거절기능을 적용한 핵심어 인식시스템

(1) 기본 핵심어 인식시스템

전체시스템은 핵심어 검출부분과 검출된 핵심어를 검증하는 두부분으로 구성된다. 핵심어 검출부에서 핵심어 모델은 실제 해당 시스템이 인식하고자 하는 핵심어만으로 구성되고, 필러 모델은 화자의 발성 중에서 핵심어 구간을 제외한 나머지 부분을 흡수하기 위해서 핵심어를 제외한 음성부나 비음성부로 구성된다. 필러 모델은 실제 핵심어 검출기의 성능향상에 큰 영향을 미치는 것으로서 묵음모델(Silence 모델), Hesitation 모델, 핵심어 이외의 단어모델, Noise 모델, 비경음향 모델 등이 이에 포함된다.

검출된 핵심어를 검증하는 부분은 핵심어의 트라이폰 열에 대한 Anti-model의 신뢰도 값을 결정하여 미리 정해둔 임계값과 비교하여 핵심어를 인정할 것인지, 거절할 것인지 결정한다.



(2) Anti-Model 모델링

GMM(Gaussian Mixture Model)은 출력확률밀도함수가 가우시안 밀도혼합(Gaussian density mixture)인 1개의 상태만으로 구성된 CHMM(Continuous HMM)의 한 형태이다.

첫째, GMM은 음향학적클래스(Acoustic Class)의 집합을 모델링할 수 있다. 화자의 발성음에 대응되는 음향공간은 모음이나 비음, 파찰음과 같은 음소를 표현하는 음향학적클래스의 집합으로 표현할 수 있는데, 이러한 음향학적클래스는 화자의 성도에 대한 정보를 가지고 있다. i^{th} 음향학적클래스의 스펙트럼 형태는 i^{th} component 밀도의 평균 μ_i로 표현되고, 평균 스펙트럼 형태의 변화는 공 분산행렬 Σ_i로 표현된다. 모든 학습 및 테스트의 음성은 레이블 되지 않기 때문에, 음향학적클래스는 은익(hidden)으로 볼 수 있다. 독립특징벡터를 가정하면, 이러한 hidden 음향학적클래스로부터 추출된 특징벡터의 관측밀도가 Gaussian mixture이다.

둘째, Gaussian basis함수의 선형조합은 샘플분포(sample distribution)의 클래스를 표현할 수 있다는 것이다. GMM의 성질 중 하나가 임의의 형태를 가지는 밀도를 부드러운 형태로 근사시키는 것이다. unimodal 가우시안 음소모델은 평균벡터(mean vector)와 공분산(covariance)으로 각 음소의 특징벡터의 이산집합으로 음소분포를 표현한다. 이와 같은 점을 고려하여 구성된 GMM은 가우시안 함수의 이산집합을 사용하여, 각각의 평균과 공분산을 가지게 함으로써 이들 두 모델의 특징을 혼합한 형태이다.

사용된 GMM의 공식은 다음과 같다.

$$P(x) = \frac{1}{\sqrt{det\sum (2\pi)^d}} e^{-\frac{(x-\mu)^T \sum^{-1}(x-\mu)}{2}} \quad (2.1)$$

이 때 d는 특징 파라미터의 차수를 나타내고, μ는 가우시안 모델의 평균을, $\sum$는 가우시안 모델의 공분산 메트릭스를 나타낸다.

(3) Anti-Phoneme Model

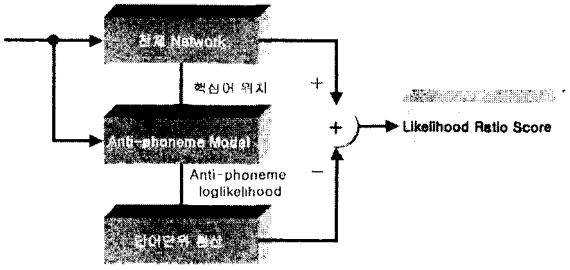
Anti-phoneme(반음소) 모델은 자기 음소를 제외한 유사음소 집합을 말하는데 일반적으로 유사음소 집합이 많을수록 반 음소가 잘 모델링 되지만, 유사 음소 집합의 크기가 너무 크게 되면 훈련 데이터량이 너무 많아지는 단점이 있다.

잘 훈련된 음소 모델만 있으면 특별한 훈련을 거치지 않고 anti-phoneme model 모델을 만들 수 있다. 또한 유사음소 집합도 자기 음소와 묵음을 제외한 나머지 모든 음소를 포함하는 52개의 음소들을 mono-phone으로 구성한 다음 tri-phone으로 구성된 자기음소를 제거한 52개의 유사음소 모델을 구성한 다음 음소 네트웍을 구성하여 핵심어 인식 음소 시퀀스 열이 매칭 되도록 하였다. 그리고 검증 오류의 최소화를 위해 자기 음소를 제외한 나머지 음소(묵음제외)들의 best Gaussian, 2nd best Gaussian, 3nd best Gaussian의 가중치, 평균, 분산을 취한다. 천이 확률은 자기 음소를 제외한 나머지 음소들의 평균을 구하여 사용한다. Gaussian mixture의 수에 따라 실험을 실시하여 거절 성능을 평가하여 최적의 Gaussian mixture의 수를 결정하여 최적의 모델을 구성하였다.

(4) 검출된 핵심어의 신뢰도

Viterbi 탐색을 사용함으로, 인식된 단어는 내부적으로

음소단위로 인식이 된다. 따라서 인식된 음소단위들을 각각의 반음소 모델과 비교하여 신뢰도를 구하고, 음소단위의 신뢰도를 단어단위의 신뢰도로 환산하기 위해서 음소단위 신뢰도의 평균을 사용한다.



패턴 (즉, $l=i(q)$)인 일반적인 음소에 대해서

$$g_{i(q)}(O_q) = \log p(O_q|\Theta_l^{(k)}) \quad (2.2)$$

$$G_{i(q)}(O_q) = \log p(O_q|\Theta_l^{(a)}) \quad (2.3)$$

이 적용되며, (2.2),(2.3)를 계산하기 위해 관측확률

$$b_j(O) = \max_{1\le k\le (mixture N\times 52)} \{ c_{jk} N(o, \mu_{jk}, U_{jk}) \} \quad (2.4)$$

과 같이 각 상태에서 최대 score를 내는 가지만을 사용한다. 여기서, c_{jk} 는 가지의 가중치이며 j는 각 음소의 상태를 뜻하며, k는 각 상태의 가지를 뜻한다.
$N(\cdot)$는 각 가지의 가우시안 분포를 의미하며, μ_{jk}는 평균 벡터, U_{jk}는 covariance matrix 이다.
여기서 $Lr_{i(q)}(O_q;\Theta)$는

$$Lr_{i(q)}(O_q;\Theta) = \frac{g_{i(q)}(o_q) - G_{i(q)}(o_q)}{|g_{i(q)}(o_q)|} \quad (2.5)$$

와 같이 정의하여질 수 있다.
식 (2.5)은 null hypothesis, $g_{i(q)}(o_q)$와 alternative hypothesis, $G_{i(q)}(o_q)$의 비를 이용하여 음소단위의 신뢰도를 개선하였다고 말할 수 있다.

여기서 음소단위의 신뢰도는 null hypothesis, $|g_{i(q)}(o_q)|$으로 정규화 함으로써 프레임 길이로 정규화하는 것 보다 음성 인식에 사용되는 문법의 변화에 탄력있고, 일관적인 음소단위의 검증성능을 보이게 된다.

다음과 같이, N개의 서로 다른 패턴, 즉 $\Theta=\{\Theta_1,\ldots,\Theta_l,\ldots,\Theta_N\}$에 상응하는 발화 검증 모델을 사용하는 신뢰도를 선택한다. 각 패턴 l 에 대해서, 음소 모델을 $\Theta_l^{(k)}$라 표시하고, anti-model인 반음소 모델을 $\Theta_l^{(a)}$라 표시한다. (즉, $\Theta_l=\{\Theta_l^{(k)},\Theta_l^{(a)}\}$)
따라서, 음소 단위들을 평균 낸 단어단위의 신뢰도는

$$s_l(O;\Theta) = \log\left[\frac{1}{N(i)}\sum_{q=1}^{N(i)} \exp\{f\cdot Lr_{i(q)}(O_q;\Theta)\}\right]^{\frac{1}{f}} \quad (2.6)$$

와 같이 되며, 이때 신뢰도가 임계값 τ_k 이하라면 거절하게 된다.

(5) 거절기능의 계산량 감소

가) 핵심어 분석을 통한 음소 제한

먼저 음소 수를 줄이기 위해 PLU(phoneme likely unit) 셋의 음소 수를 줄이는 방법이 가능하다. 그러나 그보다 핵심어의 초성과 중성을 분석하여 음소 매칭을 하여 그 외의 음소가 핵심어로 인식되었을 경우 오인식으로 간주하는 방법을 생각해 본다. 총 552개의 핵심어의 초성과 중성을 분석한다.

표 2-1 분석결과

	핵심어 음소
초성에 올 수 있는 음소(15개)	ㄱ, ㄲ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ, ㅋ, ㅌ, ㅍ, ㅎ
중성 중성에 올 수 있는 음소(14개)	ㅏ, ㅑ, ㅓ, ㅕ, ㅖ, ㅐ, ㅗ, ㅛ, ㅜ, ㅠ, ㅣ, ㅢ, ㅡ, ㅘ

52개의 총 음소중에서 15개의 초성과 14개의 중성안에 핵심어의 초성과 중성이 매칭되지 않는다면 비핵심어로 간주하여 거절시킨다. 그리고 Antiphoneme-model의 네트웍 구성에서 초성과 중성에 대하여 제한을 가하므로 계산 시간을 줄일 수 있다.

나) 음성학적 지식을 이용한 유사음소 clustering

발성기관에 의한 조음 방법 및 조음 위치 등의 음성학적 지식을 기반으로 하여 유사한 성질을 지닌 음소들을 통합하는 방법이다.

다음 표는 음성학적 지식을 이용한 모델링 그룹이다.

표 2-2 단음소의 음성학적 모델링

구 분	음 소
파열음	ㅂ, ㅃ, ㅍ, ㄷ, ㄸ, ㅌ, ㄱ, ㄲ, ㅋ
마찰음	ㅅ, ㅆ, ㅎ
파찰음	ㅈ, ㅉ, ㅊ
비음, 유음, 유성자음	ㄴ, ㅁ, ㅇ, ㄹ
모음	ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, ㅛ, ㅜ, ㅠ, ㅡ, ㅣ

다) 통계적 지식을 이용한 유사음소 clustering

Monophone모델의 확률 분포로부터 모델들끼리의 거리를 구할 때 Weighted Euclidean dostance를 사용하여 거리를 구하고, modifie K-mine(MKM)알고리즘에 의해서 monophone 모델들을 몇 개의 그룹으로 나누는 방식이다.
이때 음소 모델들 사이의 거리 척도는 다음과같이 주어진다.

$$D_{WE}(p_i, p_j) = \sum_{s=1}^{N} D_s(p_i, p_j) \quad (5.1)$$

여기서 p_i, p_j 는 각각 I 와 j 번째 음소를 나타내고, N 은 음소모델의 상태수를 나타낸다. $D_s(p_i, p_j)$는 상태간의 거리로서 다음과 같이 주어진다.

$$D_s(p_i, p_j) = \frac{1}{2} \sum_{d=1}^{v} \frac{(\mu_{isd} - \mu_{jsd})^2}{\sigma_{isd} - \sigma_{jsd}} \quad (4.2)$$

여기서 V는 음성 특징벡터의 차원이고 μ_{isd}, μ_{jsd}, σ_{isd}, σ_{jsd} 는 각각 i번째 및 j 번째 음소의 s 번째 상태의 d 차원의 평균 및 표준 편차이다. 이 방법을 이용하여 유사음소들의 군집화를 시킨 다음 코드북을 형성하여 매칭시간을 줄인다.

3. 실험 및 고찰

(1) 음성 BD

본 논문에서 사용하는 DB는 전국 대학구내 전화안내망을 목표로 하였다. 대학구내 전화안내망을 위한 핵심어 검출기를 위해서 핵심어 552개와 비핵심어 166개를 선정하여 실험하였다. 현장감을 높이기 위해서 실제 구내교환소에서 전화안내를 채록하여 녹음문장을 선정하여 훈련 데이터베이스와 테스트 데이터베이스를 구축하였다. 훈련 데이터베이스는 성인남성 30명을 대상으로 하여 1301 문장으로 구축되었으며, 테스트 데이터베이스는 성인남성 1명을 대상으로 하여 100 문장으로 구축되었다.

(2) 성능평가

표 3-1 거절기능 실험 결과

항목 / mixture수	CA	FAI	FR	CR	FAO
GMM mixture 1	70.89	5.06	24.05	67.50	32.50
GMM mixture 2	69.51	8.54	21.95	77.5	22.5
GMM mixture 3	78.04	8.54	13.41	72.5	27.5

표 4-2 계산량에 따른 거절기능 실험 결과(mixture3)

항목 / 방법	CA	FAI	FR	CR	FAO
음소제한을 가한경우	75.89	5.03	19.02	72.50	27.50
음성학적 지식을 이용한 경우	73.60	8.48	17.92	78.50	21.50
통계적 지식을 이용한 경우	78.02	8.49	13.48	79.50	20.50

4. 결론

본 논문에서는 GMM mixture수를 늘려 가면서 실험을 실시하여 mixture수를 추가하면 유사음소 개수가 늘어나게 되어 성능이 향상됨을 보아 주었다. 그러나 앞에서 언급하였듯이 유사음소의 수가 많아지면 탐색 시간이 늘어나 계산량이 기하 급수적으로 늘어남을 보이게 됨으로 계산량을 줄이기 위한 방법으로 세가지를 선택하여 실험을 한결과 통계적인 방법을 이용한 결과가 가장 좋았으며, 그 결과 인식 성능에는 변화가 없이 계산량을 줄이게 되었다.

향후에는 mixture수 더욱 증가시키는 상황에서 clustering을 실시하여 최적의 mixture수와 clustering에 의한 음소집합 코드북을 구성하여 실시간 처리가 가능한 시스템을 구성할 계획이다.

참고문헌

[1] P. Jeanrenaud, K. Ng, M. Siu, etc. Phonetic-based word spotter : Various configurations and application to event spotting", Proc. ESCA eurospeech, 1993.

[2] GuoQing, Yan Yonghong, etc, "Keyword spotting in auto-attendant system", Proc. ICSSP 2000.

[3] Jochen Junkawitsch, Gunther Ruske, Harald Hoge, "Efficient methods for detecting keywords in continuous speech", Proc. eurospeech 97, 1997.

[4] 김기태, 문광식, 김회린 외, "가변어휘 단어 인식에서의 미등록어 거절 알고리즘 성능 비교", 한국음향학회지 제 20권 제 2호, pp. 27~34, 2001.

[5] Takatoshi Jitsuhiro, Satoshi Takatoshi, Kiyoaki Aikawa, "Rejection of Out-of-Vocabulary Words using Phoneme Confidence Likelihood", ICSSP, pp. 217~220, 1998.

[6] Gethin Williams, Steve Renals, "Confidence Measures from Local Posterior Probablity Estimates" Department of Computer Science University of Sheffeldwords from registered Vocabulary " ICSP '97. Vol1, pp351~354 August, 1997.

[7] 김우성, 구명완 " 반음소 모델링을 이용한거절기능에 대한 연구" 한국음향학회지 18권 3호, pp.3~9

[8] 김민정, 석수영, 정현열, "Gaussian Mixture Model을 이용한 실시간 문맥독립화자인식에 대한 고찰", 한국음향학회 제 20권 제 1(s)호, pp.123~126,