미등록어 거절 알고리즘에서 음소 특성 추출의 신뢰도 측정 개선

오상엽*

Reliability measure improvement of Phoneme character extract In Out-of-Vocabulary Rejection Algorithm

Sang Yeob, oh*

요 약 통신 모바일 단말기에서 어휘 인식 시스템은 부정확한 어휘로부터 음소 특징을 추출하기 때문에 음소를 인식하지 못하거나 유사한 음소 오인식 오류로 인한 낮은 인식률의 문제점을 가진다. 이러한 문제를 해결하기 위해서, 본 논문에서는 입력 음소는 음소 유사율 처리를 통해 음소 사이의 거리를 측정하여 수치로 나타내고, 신뢰도 측정을 통하여 인식되어진 결과를 확인하는 시스템을 제안하였다. 이로 인해 부정확한 어휘 제공으로 인한 오인식 오류를 최소화하였으며 음소 유사율과 신뢰도를 이용하여 오류 보정율을 구하였다. 기존 방법인 에러 패턴 학습을 이용한 시스템과 의미기반을 이용한 시스템의 성능 평가 결과 2.7%의 인식 향상율을 보였다.

주제어 : 미등록어 거절, 음소 신뢰도 측정, 음소 유사율

Abstract In the communication mobile terminal, Vocabulary recognition system has low recognition rates, because this problems are due to phoneme feature extract from inaccurate vocabulary. Therefore they are not recognize the phoneme and similar phoneme misunderstanding error. To solve this problem, this paper propose the system model, which based on the two step process. First, input phoneme is represent by number which measure the distance of phonemes through phoneme likelihood process. next step is recognize the result through the reliability measure. By this process, we minimize the phoneme misunderstanding error caused by inaccurate vocabulary and perform error correction rate for error provrd vocabulary using phoneme likelihood and reliability. System performance comparison as a result of recognition improve represent 2.7% by method using error pattern learning and semantic pattern.

Key Words : Out-of-Vocabulary rejection, Reliability measure, Phoneme Likelihood

1. 서론

어휘 인식 기술 발달과 정보 통신 모바일 기기의 발전으로 어휘 기반 검색 시스템, 자동 응답 시스템 등 어휘 인식을 인터페이스로 하는 시스템들이 개발되고 있다. 하지만 어휘 인식 시스템에서 어휘 인식은 홍채 인식, 지문 인식과 달리 여전히 둔화된 발전 속도를 보이기 있다. 어휘 인식의 속도를 둔화시키는 가장 큰 원인은 부정확한 어휘 입력으로부터 음소 단위의 모델을 구성하여 인식 시 사용하므로 유사한 음소로의 인식과 다른 어휘로의 인식인 오인식을 들 수 있다. 이를 보완하기 위한 방법으로 신호 처리 단계에서의 어휘 인식 오류 보정에 대한 여러 가지 연구가 진행되고 있다[1].

시스템의 발전으로 인해 어휘 인식 시스템은 화자 독립과 화자 종속으로 여러 형태로 발전하고 있으며, 어휘 후처리에서 오류 보정에 대한 연구가 진행되고 있다[2].

오류 보정 방법은 단순한 언어 모델이 가지는 한계점을 극복하지 못한다는 단점과 오류 패턴 DB가 필요하며, 문장이 간결하고 사용자가 검색하고자 하는 핵심어로만

※ 이 논문은 2012년도 가천대학교 교내연구비 지원에 의한 결과임.
*가천대학교 IT 대학 인터랙티브미디어 교수(교신 저자)
논문접수: 2012년 6월 20일, 1차 수정을 거쳐 심사완료: 2012년 7월 10일

이루어진 경우가 많으므로 의미적으로 분석하기 힘들다. 이러한 단점을 보완하기 위해 정확한 어휘의 입력으로 인식 모델을 구성하고 인식 시 부정확한 어휘를 필터링하여 정확한 어휘로 인식하는 오류 보정 방법인 미등록어 거절을 이용한다. 입력되어진 음소를 음소 유사율 처리를 통해 음소 사이의 거리를 측정하여 수치로 나타내고 신뢰도 측정을 통하여 인식되어진 결과를 확인한다. 이로 인해 부정확한 어휘 제공으로 인한 오인식 오류를 최소화하였으며 음소 유사율과 신뢰도를 이용하여 오류 보정율을 구하였다. 기존 방법인 에러 패턴 학습을 이용한 시스템과 의미기반을 이용한 시스템의 성능 평가 결과 2.7%의 인식 향상율을 보였다.

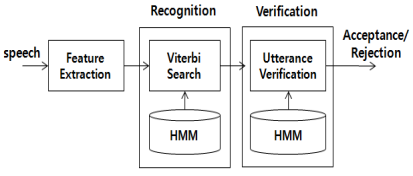
본 논문의 구성은 다음과 같다. 2장에서는 관련 연구에 대해 언급하고 3장에서는 미등록어 거절처리와 신뢰도 측정 처리에 대해 설명한다. 4장에서는 시스템 평가를 수행하고 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

2.1 마등록어 거절

음성 인식 시스템은 음성인식 기능과 검증기능이 동시에 검색되도록 구성되어진 One-pass 시스템과 인식기의 후처리 방식으로 검증 기능을 구현하는 Two-pass방식으로 구성된다. Two-pass 방식은 기존 시스템을 그대로 적용하고 검증 과정을 추가한 것으로 구현이 쉽다는 장점을 가지고 있다. 발화 검증 시스템을 설계할 때 미등록어와 잘못 인식된 단어를 잘 선별할 수 있는 검증 모델에 기반한 적정한 신뢰도(confidence measure)를 정의해야 하고, 훈련 데이터에서 검증 오류를 최소화할 수 있도록 검증 모델을 적응시키는 훈련과정을 선택해야 하며 유사도의 변화와 검증 문턱치의 변화, 훈련과 테스트 상태의 변화에 강해야 한다.

[그림 1]은 인식과 검증으로 구성된 2단계 발화 검증 시스템의 기본 구조를 나타낸다.



[그림 1] 주어진 시간 이내의 프레임 수신 성공률

1단계에서 인식 모델을 사용하여 비터비(viterbi)탐색 알고리즘에 의한 인식과정을 수행한다. 음소 모델들은 ML(Maximum Likelihood)를 이용하여 HMM(Hidden Markov Model)의 파라미터를 최적화한다. 인식 과정을 거치는 동안 각 단어의 발화는 음소 가설로 분할되며, 그 결과를 발화 검증 시스템으로 전달한다. 두 번째 단계인 발화 검증 과정은 인식된 후보 단어의 음소열에 대해 반음소 모델과의 신뢰도를 구하여 그 단어의 신뢰도 값을 결정한다. 이 신뢰도 값이 미리 정해둔 문턱치보다 크면 등록단어로 인식이 되고 아니면 거절된다.

2.2 음소특성 추출

특성 추출은 인식에 필요한 특징 성분을 신호로부터 나타내는 과정이며 일반적으로 정보의 압축과 차원의 감소를 표현한다. 추출된 특성의 좋고 나쁨은 인식율로 판단되며 특성 추출 과정에서 청각 특성을 반영한 것으로 음향학적 모델을 이용한 필터 뱅크 분석, 주파수에 따른 대역폭의 증가, 프리엠퍼시스 필터 등이 사용된다. 어휘 신호의 동적 특성을 반영하기 위한 방법으로 켑스트럼 1차, 2차 미분 값을 사용한다[3].

단구간 대수 에너지는 프레임의 에너지를 구하는 것으로 무음 구간의 에너지 값이 음성구간의 값보다 적기 때문에, 무음 구간과 음성구간 사이의 큰 에너지 변화를 이용하여 끝점을 검출하는데 널리 사용된다. 프레임의 에너지를 E_l라 할 때, 식은 다음과 같다.

$$E_l = 10\log[\sum_{n=0}^{n-1} x^2(n)] \quad (1)$$

$x(n)$은 양자화된 신호, N은 한 프레임내의 샘플 수를 의미한다. 단구간 영 교차율은 각 프레임 내에서 수평축을 기준으로 음성 신호의 파형이 변하는 횟수를 나타낸 것으로 프레임의 유·무성음 구간을 판별하는데 사용된다. 유성음은 에너지가 낮은 주파수에 집중되어 있기 때문에 낮아지는 반면에 무성음은 에너지가 높은 주파수대역에 집중되기 때문에 높은 영 교차율을 가진다. 무음 구간에서의 영 교차율은 주위 환경에 따라 변하게 되지만 일반적으로 무성음보다는 작고 유성음보다는 큰 값을 갖게 되며 잡음환경에서 일반적으로 다음 식 (2)와 같이 표현한다.

$$Z = \frac{1}{2}\sum_{m=0}^{N-1} |\mathrm{sgn}[x(n-m)] - \mathrm{sgn}[x(n-m-1)]|,$$
$$\mathrm{sgn}[x(n)] = \begin{cases} 1, & x(n) > 0 \\ -1, & \text{otherwise} \end{cases} \quad (2)$$

Z는 영 교차율을 나타낸다. 영 교차율은 음성신호의 값에 의해 많은 영향을 받으며, 샘플링 전의 값을 정확히 결정하여야 한다. 따라서 값에 의해 레벨교차율을 구하며 검출하는 과정은 다음 식(3)과 같다.

$$Z = \frac{1}{2}\sum_{m=0}^{N-1}|sgn[x(n-m)] - sgn[x(n-m-1) - L_{th}]|,$$
$$sgn[x(n)] = \begin{cases} 1, & x(n) > 0 \\ -1, & otherwise \end{cases} \quad (3)$$

L은 레벨 교차율, L_{th}는 구해진 오프세트 레벨을 나타낸다.

미분은 시간축 방향의 필터링으로 생각할 수 있으며 시간축 방향으로의 특징 벡터를 얻는 과정이다. 인식을 위하여 주로 사용되는 특징은 MFCC와 LPC가 주로 사용된다.

MFCC는 신호를 안티 엘리어싱 필터(anti-aliasing filter)로 거친 다음, A/D 변환을 거쳐서 디지털 신호 x(n)로 변환한다. 디지털 신호는 고대역 통과 특성을 갖는 디지털 프리엠퍼시스 필터를 거친다. 프리엠퍼시스된 신호는 해밍 윈도우(hamming window)를 씌워서 블록 단위의 프레임으로 나누어지고 프레임 단위로 만들어진다. 프레임의 크기는 보통 20-30ms이며 프레임 이동은 10ms가 사용한다. 한 프레임의 신호는 FFT(Fast Fourier Transform)를 이용하여 주파수 영역으로 변환하여 주파수 대역을 여러 개의 필터뱅크로 나누고 각 뱅크에서의 에너지를 구한다. 밴드 에너지에 로그를 취한 후 DCT(Discrete Cosine Transform)를 취하면 최종적인 MFCC가 언어진다. MFCC 계수는 12개를 사용하며 이와는 별도로 구한 프레임 로그 에너지가 추가적으로 사용되어 인식의 입력으로 사용되는 특성 벡터는 13차 벡터로 구성되어 사용된다[4].

3. 시스템 모델

3.1 미등록어 거절 처리

미등록어 거절 방법에는 가변 어휘 단어 시스템에서 비터비 탐색 시 사용되는 네트워크망을 이용하며, 구성된 네트워크망에서 인식된 결과는 등록어들과 음소들의 열로 나타나게 된다. 즉, “묵음+(등록어 및 음소들의 열)+묵음”과 같은 형태가 된다. 단어 패널티를 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 나타난다. 이렇게 인식된 결과를 발화 검증 시스템으로 넘기게 되며 가변 어휘 단어 인식 시스템의 단어 패널티와 인식된 결과의 삽입된 음소들의 개수를 이용하여 미등록어를 거절시킨다.

핵심어 검출 네트워크의 핵심어 모델들과 필러 모델들은 GMM(Gaussian Mixture Model)을 이용하여 어휘 모델에 대한 형태로 모델링한다.

본 시스템에서 핵심어 모델과 필터 모델은 인식 네트워크에서 병렬로 연결되어 매 프레임마다 새로운 입력을 받아들인다. 핵심어의 오검출을 방지하기 위해 핵심어가 실제 발성 되었는지를 검증하고 신뢰도를 계산한다.

삽입된 음소들은 필러 모델을 뜻하며 삽입된 음소가 많다는 것은 인식 결과에 핵심어가 없다는 의미이다[5].

단봉(unimodal) 가우시안 음소 모델은 평균 벡터(mean vector)와 공분산(covariance)으로 각 음소의 특징 벡터의 이산 집합으로 음소 분포를 표현한다. 이와 같은 점을 고려하여 구성된 GMM은 가우시안 함수의 이산 집합을 사용하여, 각각의 평균과 공분산을 갖는다. GMM은 1상태로 구성되는 특성 때문에 상대적으로 모음과 자음의 DB 용량차이에서 일어나는 확률 분포의 차이를 최소화 할 수 있으므로 음소를 이용한 연속 인식 네트워크에 적합한 장점을 가지고 있다.

3.2 신뢰도 측정 및 처리

인식된 결과는 음소나 단어로부터 발화되었을 확률에 대한 상대 값을 의미하며 신뢰도는 인식 결과에 대해 그 결과가 얼마나 믿을 만한 것인가를 나타내는 척도이다[6].

어떤 O를 관측 세그먼트라고 하면 인식 과정에서 O가 입력되었을 때는 두 가지의 가정이 가능하다. O가 실제 세그먼트 k 일 것이라는 가정이 가능하며 영가설이라 하고 H_0로 표시한다. 또한, O가 실제 세그먼트가 아닌 다른 유사 발화라 가정이 가능하며 대립가설이라 하고 H_1으로 표현한다. 주어진 테스트 세그먼트 O에 대해 발화 검증 과정은 영가설에 대한 확률과 대립가설에 대한 확률을 비교하여 영가설에 대한 확률이 크면 인식하고 아니면 잘 못된 인식으로 판단한다. 따라서 영가설에 대한 확

률과 대립가설에 대한 확률을 비교하기 위한 식 (4)로 표현한다.

$$P(O|H_0) > P(O|H_1) \quad (4)$$

베이시안의 정리는 불확실한 상황에서의 의사 결정 문제를 수리적으로 다룰 때 사용한다. 연역적 추론 방식인 확률을 사용하여 귀납적 추론을 만들어내는 방법이다. 전통적인 확률을 직접적인 확률(direct probability)이라 한다면 베이시안의 정리는 역의 확률(inverse probability)이라 하고 식 (5)과 같이 정리한다.

$$P(H_0|O) = \frac{P(O \cap H_0)}{P(O)} = \frac{P(O \cap H_0) \cdot P(H_0)}{\sum_{k=1}^{N} P(O \cap H_k) \cdot P(H_k)} \quad (5)$$

위의 식 (4)를 베이시안 룰(bayes rule)인 식 (5)에 의해 정리하면 식 (6)과 같이 표현되며

$$\frac{P(H_0|O)P(H_0)}{P(O)} > \frac{P(H_1|O)P(H_1)}{P(O)} \quad (6)$$

$P(H_0|O)$는 모델 λ_k에서 O가 관측될 확률이고, $P(H_1|O)$는 다른 모델에서 O 가 관측될 확률로 표현한다. 식 (6)의 $P(O)$는 결정규칙에 영향을 미치지 않는 동일한 값이므로 제거하여 정리하면 식 (7)과 같다.

$$\Lambda(O) = \frac{P(H_0|O)}{P(H_1|O)} > \frac{P(H_1)}{P(H_0)} \quad (7)$$

식 (7)로 나타내어진 $\Lambda(O)$항을 우도비 (likelihood ratio)라 하며 신뢰도 측정을 위해 사용된다. H_1을 모델링하기 위해서 각 음소마다 유사한 음소들을 구하여 파라미터로 훈련하고 훈련된 파라미터를 $\lambda_{\bar{k}}$로 표현한다. 모델 λ_k에서 관측 세그먼트 O가 관측될 확률과 훈련된 파라미터를 $\lambda_{\bar{k}}$로 표현하면 식 (8)와 같이 표현한다.

$$P(O|\lambda_k) > P(O|\lambda_{\bar{k}}) \quad (8)$$

모델 λ_k와 훈련된 파라미터 $\lambda_{\bar{k}}$ 로 표현된 형태를 안티모델로 사용하고 안티모델에 log을 취해서 우도(log-likelihood)로 표현하면 식(9)과 같다.

$$LLR_k(O, \lambda_k) = \log P(O|\lambda_k) - \log P(O|\lambda_{\bar{k}}) \quad (9)$$

$LLR_k(O, \lambda_k)$은 모델 λ_k에서 관측 세그먼트 O가 관측될 확률을 계산하는 로그 우도비 값으로 표현된다. 우도 값이 너무 큰 범위에서 나타나지 않도록 정화규하여 시그모이드 함수를 사용하고 최종적인 음소 신뢰도를 계산하게 된다.

$$f(LLR) = \log \frac{1}{1+\exp^{(-a \cdot LLR)}} \quad (10)$$

식 (10)은 로그우도비를 시그모이드 함수를 사용하여 정규화한 식을 표현한다. 신뢰도를 측정하기 위해 식 (10)에 표현한 로그우도비를 사용하여 측정하고 측정된 로그우도비가 높을수록 신뢰도가 높으며 인식 시에 정확한 인식으로 표현된다.

4. 실험 결과 및 분석

본 실험에서는 다양한 음소의 조합을 고려한 PBW 445DB를 이용하였다. 어휘수가 총 445개로 구성되어 있으며 1명이 2회 발성한 것을 1개의 set로 구성하였다. 이러한 set이 남성음이 10set, 여성음이 10set으로 모두 20set으로 구성하였다[7].

실내 환경과 잡음 환경에서 이동기기에 내장되어 있는 내장형 마이크로폰을 사용하여 16kHz Mono로 녹음하였고, 16bit PCM 양자화를 사용하였다. 녹음된 데이터는 인식기 학습을 위해 MFCC 특성 추출 방법을 사용하였고 인식기는 SITEC에서 개발한 ECHOS[8]를 이용하였다. HMM 기반 ECHOS는 각 단어별 데이터로 학습된 인식 모델을 이용하여 발화된 단어의 인식 가능한 단어들의 인식 가능도를 우도로 표현하고 최대값을 가지는 단어를 최종 결과로 선정한다. 미등록어 거절의 성능은 다음과 같은 항목을 기준으로 평가하였다[9].

1) 등록어

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

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

③ FR(False Rejected for Keyword)
인식 대상 등록어를 말했는데 reject한 경우의 확률
④ CA + FAI + FR = 100%

2) 미등록어

① CR(Correctly Rejected for OOV)
미등록어에 대해 reject한 경우의 확률
② FAO(False Accepted Out-of-Grammar Word, OOV)
미등록어인데 accept한 경우의 확률
③ CR + FAO = 100%

어휘에 대한 인식 실험을 통하여 오류 보정에 대한 신뢰도와 보정율은 식 (10)과 다음 식 (11)을 이용하여 표 1에 나타내었다.

$$R = \lambda \frac{1}{n} \sum_{k=1}^{n} (1 - \alpha_k) b_k, \qquad \lambda = \sum_{k=1}^{n} \alpha_k \qquad (11)$$

표 1은 기존의 에러 패턴 학습을 이용한 방법[10][11]인 error pattern과 의미기반의 방법[12]인 semantic 그리고 본 논문의 제안 방법인 out-of-vocabulary rejection의 결과를 나타내었다.

〈표 1〉 오류 보정률 비교

오류 보정	인식률(%)	보정률(%)
error pattern	76.2	3.1
semantic	79.4	2.5
rejection	82.1	2.7

에러 패턴 학습을 이용한 오류 보정의 경우 3.1%, 의미 기반의 오류 보정의 경우 2.5%를 보였으며, 본 논문에서 제안한 out-of-vocabulary rejection을 이용한 음소 유사율을 사용할 경우 인식율 82.1%, 보정률 2.7%를 보였다. 인식률은 비교 시스템에 비해 5.9%와 2.7% 향상되었으며, 보정률 2.7%는 error pattern과 semantic의 평균값에 해당하며, 시스템의 안정적인 미등록어 거절 처리를 의미한다.

5. 결 론

본 연구에서는 미등록어 거절을 이용한 음소 유사율 처리를 통해 음소 사이의 거리를 측정하여 수치로 나타내고 신뢰도 측정을 통하여 인식되어진 결과를 확인하였다. 음소 유사율 처리는 음소 사이의 거리를 측정하여 비슷한 음소를 찾아 관리하며 신뢰도 측정 및 처리에서는 음소의 신뢰도를 측정 및 처리한다. 이로 인해 부정확한 어휘 제공으로 인한 오인식 오류를 최소화하였으며 음소 유사율과 신뢰도를 이용하여 오류 보정율을 구하였다. 기존 방법인 에러 패턴 학습을 이용한 시스템과 의미기반을 이용한 시스템의 성능 평가 결과 2.7%의 인식 향상율을 보였다.

참 고 문 헌

[1] E. K. Ringer and J. F. Allen, "A fertility channel model for post-correction of continuous speech recognition", Proc. ICSLP, pp.897-900, Oct, 1996.
[2] 박미성, 김미진, 김계성, 최재혁, 이상조, "The syllable recovrey rule-based system and the application of a morphological analysis method for the post-processing of a continuous speech recognition", 전자공학회논문지, 제36권, 제3호, 47-57쪽, 1999년 3월.
[3] M. F. Gales, "Model-based techniques for nosie robust speech recognition", Ph. D. dissertation, University of Cambridge, Sept, 1995.
[4] M. Ostendorf, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition", Speech and Audio Processing, IEEE, Vol.4, pp.360-378, 1996.
[5] T. Jitsuhiro, S. Takatoshi, and K. Aikawa, "Rejection of out-of-vocabulary words using phoneme confidence likelihood," ICASSP, pp. 217-220, 1998.
[6] 송원문, 김명원, "문맥 및 사용 패턴 정보를 이용한 음성 인식 후처리", 정보처리학회논문지, 제13-B권, 제5호, 553-560쪽, 2006년.
[7] S. Kaki, E. Sumita, and H. Iida, "A method for correction speech recognition using the statistical features of character co-occurrence", Proc. COLING-ACL, pp.653-657, Aug, 1998.
[8] 김용현, 정민화, "에러패턴 학습과 후처리 모듈을 이

용한 연속 음성 인식의 성능향상", Proc. KISS Spring Semiannual Conf. 제27권, 제1호, 441-443쪽, 2000년 4월.

[9] M. W. Jeong, B. C. Kim, and G. G. Lee, "Semantic-oriented error correction for spoken query processing", Proc. IEEE Workshop on ASRU, pp.156-161, Nov, 2003.

[10] K. Demuynck, J. Duchateau, and D. Van Compernolle, "A static lexicon network representation for cross-word context dependent phones," In Proc. EUROSPEECH, Vol.1, pp. 143-146, 1997.

[11] A. S. Manos and V. W. Zue, "A study on out-of-vocabulary word modeling for a segment-based keyword spotting system," Master Thesis, MIT, 1996.

[12] S. Ortmanns, A. Eiden, H. Ney, and N. Coenen, "Look-ahead Techniques for Fast Beam Search," InProc. IEEE ICASSP-1997, pp. 1783-1786, 1997.

오 상 엽

・저자약력
1999 : 광운대학교
전자계산학과 이학박사.
현재 : 경원대학교 IT대학
인터랙티브미디어 교수
・관심분야
소프트웨어공학, 버전관리, 소프트웨어재사용, 형상관리, 객체지향, 음성인식, 음성/음향 신호처리
・E-Mail : syoh@gachon.ac.kr