Viterbi 탐색 특성을 이용한 미등록어휘 제거에 대한 연구

김규홍, 김회린
한국정보통신대학교

A Study on OOV Rejection Using Viterbi Search Characteristics

Kyuhong Kim and Hoirin Kim
Information and Communications University

kkh@icu.ac.kr, hrkim@icu.ac.kr

Abstract

Many utterance verification (UV) algorithms have been studied to reject out-of-vocabulary (OOV) in speech recognition systems. Most of conventional confidence measures for UV algorithms are mainly based on log likelihood ratio test, but these measures take much time to evaluate the alternative hypothesis or anti-model likelihood. We propose a novel confidence measure which makes use of a momentary best scored state sequence during Viterbi search. Our approach is more efficient than conventional LRT-based algorithms because it does not need to build anti-model or to calculate the alternative hypothesis. The proposed confidence measure shows better performance in additive noise-corrupted speech as well as clean speech.

Ⅰ. 서론

최근 음성인터페이스 기술이 발전함에 따라 음성인식기의 수요가 늘어가고 있다. 현재까지의 기술수준으로 제한된 영역에서의 음성인식이나, 비교적 조용한 환경에서의 음성인식은 상당히 우수한 인식성능을 보여주고 있으며, 상용화 제품이 출시되고 있다. 하지만, 음성인터페이스에서 사용자가 등록되지 않은 어휘를 발성할 경우와 잡음의 영향으로 인하여 오인식될 경우에는 음성인식기에서 오류가 발생하게 되는데, 이 오류를 감지할 수 있어야 보다 자연스러운 인터페이스를 구현할 수 있다. 인식된 결과를 얼마나 신뢰할 수 있는지의 여부를 수치로 표현한 것을 신뢰도(confidence score)라 하며, 이 신뢰도를 기반으로 인식결과가 틀렸는지 맞았는지 검증하는 것을 발화검증(utterance verification)이라 한다. 음성인식기에서는 사용자가 미등록 어휘를 발성하였을 경우, 음성인식기에서는 등록된 어휘들 중 인식스코어가 가장 높은값을 나타내는 단어를 인식결과로 채택한다[1]. 이때 음성인식기에서는 음향사전확률(acoustic prior probability)이 무시된 최대우도(maximum likelihood)를 기반으로 인식스코어를 계산하기에 이 스코어 자체가 발화검증에 사용될 수 없다[2]. 대부분의 발화검증방법은 LRT(Likelihood Ratio Test)를 기반으로 하여 신뢰도를 계산하고, 이를 미리 정의한 임계치와 비교하여 거절여부를 결정한다. 따라서, LRT기반의 발화검증 알고리즘은 인식과정에서 무시된 음향확률을 최대한 정확하게 추정하면서도 계산량이 크지 않아야 한다.

Viterbi 탐색기에서는 최고의 스코어를 갖는 탐색 경로를 추적하는데, 이외의 경로는 인식결과에 반영하지 않는다. 본 논문에서는 매 프레임마다 Viterbi 탐색 중에 나타나는 누적 탐색 스코어를 이용하여 최대의 스코어

를 갖는 마지막 스테이트를 계속 추적하고, 이를 발화검증에 사용하고자 한다.

제안된 발화검증법은 alternative hypothesis나 antimodel에 대한 관측우도(observation likelihood)를 계산할 필요가 없기 때문에 계산이 간단하다는 특징이 있다.

Ⅱ. LRT기반 신뢰도값

1. Baseline 시스템

본 논문에서 사용한 음성인식기는 CHMM (Continuous Hidden Markov Model) 기반의 고립 단어 인식기이며, 음성인식의 결과를 발화검증기에서 후처리 형식으로 신뢰도를 측정한다. 발화검증의 구현의 측면에 있어서 크게 두 부류로 나눌 수 있다. 첫째 방법은 본 논문에서와 같이 후처리 방식으로 구현된 방법이고, 다른 하나는 음성인식기에 신뢰도를 반영하여 인식스코어를 그대로 발화검증에 사용할 수 있는 접근방식이 있다. 후자의 대표적인 예로, Viterbi 탐색중에 스코어에 신뢰도를 탐색스코어에 포함하여, 오인식 가능성있는 어휘들을 인식도중에 효율적으로 프루닝(pruning)할 수 있는 방식의 접근방법이다[3]. 이러한 방식은 2-stage구조에 비하여 성능저하를 최소화하면서 서비스 응답속도와 관련이 있는 turn-around time을 줄여줄 수 있는 접근방식이다. 하지만 모든 탐색 공간에서 신뢰도를 계산해야 하기 때문에 계산량이 많다는 단점으로 인하여 계산 성능이 좋은 서버환경에는 적합하나 이동전화 또는 PDA등과 같은 모바일 환경에서는 적용하기가 어렵다.

2. LRT기반의 신뢰도 측정 방법

발화검증 단계에서는 인식된 단어에 대한 신뢰도를 측정하고, 이를 미리 정의한 임계치와 비교하여 인식된 결과에 대한 거절 여부를 판단한다. 이에 대한 많은 연구가 진행되어 왔는데, 대부분의 경우가 LRT기반의 접근 방식이 주류를 이루었다. 이러한 LRT기반의 접근방식에서는 alternative hypothesis의 모델링 방법에 따라서 여러 종류의 신뢰도 측정법으로 분류된다. 결국 음향사전확률(acoustic prior probability)을 어떻게 모델링하느냐 background모델, 또는 garbage 모델[5], antimodel[3,6,7]등의 여러 방법으로 나뉘게 된다. 본 논문에서는 제안된 방법과의 성능비교를 목적으로 일반화된 신뢰도를 발화검증을 위하여 채택하였다[3]. 본 논문에서는 인식결과를 단어단위로 검증하고 forced alignment를 통하여 word를 구성하고 있는 state들의 경계 정보를 이미 알고 있다고 가정한다. x_n 을 인식된 state s_n에 해당하는 관측 벡터열이라 한다면 프레임단위의 likelihood ratio는 식 (1)로 표현된다

$$CM_{frame}(n) = log\frac{L[x_n|s_n]}{L[x_n|s_n^a]} \quad (1)$$

인식된 결과에서 n 프레임의 state가 s_n일때의 관측 특징벡터 x_n의 프레임 likelihood를 $L[x_n|s_n]$라 정의하며 s_n^a은 state s_n에 대한 antimodel이다. 프레임단위의 likelihood ratio의 분포 범위가 크기 때문에 sigmoid 함수를 적용하여 dynamic range를 줄여서 식 (2)와 같이 음소 단위의 신뢰도를 계산한다.

$$CM_{phone}(ph) = \frac{1}{\tau}\sum_{n=1}^{\tau} log[\frac{1}{1+exp(-\alpha \cdot CM_{frame}(n)+\beta)}] \quad (2)$$

식 (2)는 음소단위의 신뢰도이며, τ는 음소를 구성하고 있는 프레임수이다. α와 β값은 sigmoid 함수의 분포특성을 변화시키는 상수이다. 단어단위의 신뢰도도 마찬가지로 sigmoid 함수를 이용하여 dynamic range를 줄인 평균값을 발화검증을 위한 신뢰도로 사용하였다 다음 식 (3)은 단어단위의 신뢰도의 정의이다.

$$CM_{word}(W) = \frac{1}{N}\sum_{n=1}^{N} log[\frac{1}{1+exp(-\alpha \cdot CM_{phone}(n)+\beta)}] \quad (3)$$

윗 식에서 N은 단어를 구성하고 있는 음소의 개수이며, α와 β값은 분포특성을 변화시키는 상수이다 발화검증 방식은 이미 기술한 바와 같이 후 처리방식으로 구현하였는데, 인식된 단어구간에 대하여 식 (3)을 이용하여 신뢰도를 계산하고, 이 값을 미리 정한 임계값과 비교하여 작으면 거절하고 크면 수락하는 방법으로 구현되었다.

Ⅲ. Momentary best-scored state을 이용한 OOV 검출

Viterbi 탐색중에는 매 프레임마다 각각 state에서 Viterbi 스코어를 최대로 만들어주는 마지막 스테이트의 경계정보를 저장한다. Viterbi 탐색 도중에는 어떤 스테이트 열이 최종 인식된 state 시퀀스가 될지 알 수 없다. 하지만, 매 시간마다 현재까지 관측된 특징벡터에 가장 적합한 마지막 스테이트는 추적 할 수 있으며, 이를 MBS(Momentary Best-scored State)이라 정의한다. 만약, 인식결과가 오류없이 정확하게 인식됐다면, MBS 시퀀스와 인식결과의 state 시퀀스가 매 프레임마다 일치할 가능성이 크다. 이 말은 다른 단어들과 인식도중에 인식결과와 크게 경쟁하지 않았다는 말이되며, 오인식된 경우와 OOV일 경우에는 매 프레임마다 추적된 두가지 state 시퀀스가 프레임단위로 일치하지 않을 경우가 많다. 따라서, 이 두 가지 state 시퀀스의 유사도

가 인식기의 혼란 정도를 표현해준다고 할 수 있다. 본 논문에서는 이러한 스테이트의 일치여부를 기반으로 인식된 단어의 신뢰도를 측정하는 방법을 제안한다.

그림 1은 Viterbi 탐색 알고리즘에서 MBS를 찾는 방법을 보여준다. 주어진 관측데이터에 대한 모든 스테이트(N_{state})에서의 Viterbi score, $\delta_t(j)$가 계산되고 이들 중 최대가 되는 스테이트를 찾아내게 되는데 이를 MBS, m_t,라 정의한다.

Initialization

$$\delta_1(i) = \pi_i b_i(x_1), 1 \le i \le N_{state}$$

$$\psi_1(i) = 0, 1 \le i \le N_{state}$$

$$m_1 = \arg\max_{1 \le i \le N_{state}} \delta_1(i)$$

Recursive step

$$\delta_t(j) = \max_{1 \le i \le N_{state}} (\delta_{t-1}(i) a_{ij} b_j(x_t)), 1 \le j \le N_{state}$$

$$\psi_t(j) = \arg\max_{1 \le i \le N_{state}} (\delta_{t-1}(i) a_{ij}), \ 1 \le j \le N_{state}$$

$$m_t = \arg\max_{1 \le i \le N_{state}} \delta_t(i)$$

Final step

$$p = \max_{1 \le i \le N_{state}} \delta_T(i)$$

$$s_T = \arg\max_{1 \le i \le N_{state}} \delta_T(i)$$

$$m_T = s_T$$

Backtracking

$$s_t = \psi_t(s_{t+1}) \quad t = T-1, T-2, \ldots, 1$$

그림 1. momentary best state를 동시에 추적하는 Viterbi 탐색 알고리즘

Viterbi 탐색에서 검색된 인식결과인 state 시퀀스 ($S = s_1 s_2 \ldots s_T$) 와 MBS 시퀀스($M = m_1 m_2 \ldots m_T$)를 추척하였고, 식 (4)에서처럼 프레임 단위로 두 state 열을 비교하여 신뢰도로 나타내었다.

$$CM(S,M) = \frac{1}{T}\sum_{t=1}^{T} d(tri(s_t), tri(m_t)) \qquad (4)$$

$$d(tri_a, tri_b) = \begin{cases} 1, & \text{if } tri_a = tri_b \\ 0, & otherwises \end{cases} \qquad (5)$$

여기서, tri(s)는 state s를 포함하는 트라이폰으로 정의되며, $d(phoneme1, phoneme2)$는 두 모델간의 거리를 표현하는 거리함수이다. 본 논문에서는 식 (5)에 표현한 것과 같이 두 트라이폰이 일치하면 1 다르면 0을 나타내는 함수로 정의하였다.

기존의 LRT기반의 방법에 비하여 프레임 단위의 계산량을 비교하였을 때, 기존의 LRT기반의 신뢰도 측정방법은 likelihood 계산 2번과 sigmoid 정규화 1회의 계산이 필요한데 반하여 제안된 방법은 한번의 단순한 비교연산이 필요하므로, 제안된 방법의 계산량이 적다고 할 수 있다. 제안된 방법의 두 번째 장점은 잡음환경에서도 강인한 특성을 보여준다는 것이다. 기존의 LRT방식의 신뢰도 측정방식은 잡음이 첨가되었을 경우에는 likelihood값의 변화가 커져서 신뢰도를 계산할 때 큰 영향을 주지만, 제안된 방법의 신뢰도 측정방식은 누적확률의 순서정보를 이용하기 때문에, 잡음에 강인한 특성이 있다.

Ⅳ. 실험 및 결과

본 논문에서 실험을 위하여 사용한 데이터 베이스는 한국어 PBW(Phonetically Ballanced Word) 452DB와 POW 3848DB를 사용하였다. PBW는 72명의 남성과 여성화자가 452가지의 단어를 두 번씩 발성하여16비트로 양자화하고 16 kHz 샘플링하여 저장된 데이터베이스이다. 첫 번째 200단어 인식실험에서는 음향 모델 훈련에 PBW 데이터베이스의 약 90%를 사용하였고 나머지 10%를 인식테스트에 사용하였다. 테스트 단어의 전체 수는 7,232 개이며, 이중 등록된 단어에 해당하는 수는 3,200개 이고, 나머지 4,032개는 미등록실험에 사용되었다.

성능을 측정의 도구로 FA(False Alarm)과 FR(False Rejection)을 사용하였다. FA는 인식된 단어가 틀렸음에도 불구하고 발화검증단계에서 계산한 신뢰도가 미리 정의한 임계치보다 커서 인식결과가 맞다고 판단하는 에러이고, FR은 인식된 결과가 맞았는데도 신뢰도가 임계치보다 낮아서 거절한 경우의 에러이다. 이 두 가지 에러는 임계치의 값에 따라서 서로 반비례하며 변하게 되는데, 임계치의 변화에 따라서 각각의 에러를 그래프로 표현한 것을 ROC (Receiver Operational Characteristic) 곡선이라 한다. 이 ROC 곡선은 발화검증의 동작 특성을 모든 가능한 동작영역에서 갈 보여주기 때문에 성능평가 지표로 사용된다. 한편, 이때 FA와 FR이 같을때의 오류율을 EER(Equal Error Rate)이라 하며, 이 또한 발화검증의 성능을 표현하는 지표로 사용된다.

그림 2는 기존의 LRT기반의 발화검증방식과 제안된 방법의 성능을 비교하기위한 ROC 곡선을 보여준다. 그림에서 보여주듯이 제안된 방법의 성능이 전 동작영역에서 기존의 방법에 비하여 낮은 FA와 FR을 보여줬

다 또한 LRT기반 신뢰도 측정방식은 8.8%의 EER을 보여줬으나, 제안된 방법은 3.7%의 EER을 보여줘서 기존의 방법에 비하여 계산량에서 뿐만아니라 성능면에서도 상당히 우수한 결과를 보여주고 있다.

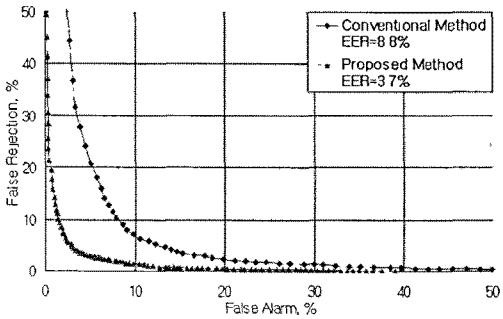

그림 2 제안된 발화검증법의 ROC 곡선

두 번째 실험에서는 인식단어의 수를 4,048개로 확장했으며, POW를 사용하여 음향모델을 훈련시키고 PBW를 사용하여 테스트에 사용하였다. 잡음이 없는 상태에서 등록된 단어 28,800개에 대한 인식률은 97.35%였으며, 발화검증을 위하여 잡음환경에서의 성능의 변화를 보기 위하여, Aurora2[4] 잡음의 일부를 PBW에 신호대잡음비가 20dB, 10dB가 되도록 잡음을 첨가하여, 비교적 조용한 잡음 환경에서의 성능을 측정하였다. 다음 표 1은 3가지 잡음에 대한 성능을 측정하였으며, 모든 경우에서 제안된 방법이 더 우수한 성능을 보여줬다. 표 2에서는 잡음이 없는 경우의 성능과 비교적 조용한 환경을 고려한 7가지 환경에서의 평균성능을 표현하였다. 각각의 경우에서도 모두 제안된 방법이 성능이 좋았다.

표 1. 다양한 잡음 환경에서의 EER (%)

Noise type	Restaurant		Babble		Car	
	20dB	10dB	20dB	10dB	20dB	10dB
LRT	39.7	42.9	40.0	42.2	30.8	39.6
Proposed	23.1	28.5	23.1	27.6	20.7	31.1

표 2 잡음환경과 환경에서의 EER(%)

Noisy condition	Clean	Multi-condition
LRT	29.4	36.8
Proposed	19.1	24.1

V. 결론

본 논문에서는 Viterbi 탐색의 특성을 이용하여 음성인식의 신뢰도값을 계산해내는 방법을 제안하였다. 본 논문에서 제안하는 방식을 발화검증에 적용하였을때 기존의 LRT방식에 비하여 단위 프레임당 계산량이 LRT 방식에 비하여 매우 적다는 특징이 있으며, 확률값의 비율을 사용한 것이 아니라 순위정보를 이용하므로 비교적 조용한 잡음 환경에서는 LRT방식 보다는 잡음에 강인한 특성을 보여주고 있다. 이러한 우수한 동작 특성으로 인하여 계산량이 중요시되는 PDA나 이동단말기 등의 모바일 환경에서 음성인터페이스에 적용되기에 적합하다.

참고문헌

[1] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing, Prentice Hall PTR, 2001.

[2] M.GHULAM, T.SATO, T.FUKUDA, and T. NITTA, "Confidence Scoring for Accurate HMM-based Speech Recognition by Using Monophone-Level Normalization Based on Subspace Method", IEICE Trans. Information and System Vol. E86-D, No 3, pp.430-437, March 2003.

[3] M. W. Koo, C. H. Lee and B. H. Juang, "Speech Recognition and Utterance Verification Based on a Generalized Confidence Score", IEEE Trans. on Speech and Audio Processing, 9(8), pp 821-832, 2001.

[4] D. Pearce,H. and G. Hirsch,"The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", in Proc. ICSLP, Beijing, China, vol 4, pp.29-32, 2000.

[5] E. Lleida and R.C.Rose, "Efficient decoding and training procedures for utterance verification in continuous speech recognition", in Proc. ICASSP'96, pp.507-510, 1996.

[6] R.A.Sukkar and C. H. Lee, "Vocabulary independent utterance verification for nonkeyword rjection in subword based speech recognition", IEEE Trans. on Speech and Audio Processing, vol. 4, no. 6, pp. 420-429, 1996

[7] M. G. Rahim, C.H.Lee, and B.H. Juang, "Discriminative utterance verification for connected digit recognition", IEEE Trans. on Speech and Audio Processing, vol 5, no.3, 1997.