1998년도 한국음향학회 학술발표대회 논문집 제 17 권 제 2(s) 호

HMM의 상태별 가중치를 이용한 핵심어 검출의 성능 향상

최동진°, 윤영선, 윤성진, 오영환
한국과학기술원 전산학과

Performance Improvement of Word Spotting Using State Weighting of HMM

Dong-jin Choi°, Youngsun Yun, Seong Jin Yun, Yung Hwan Oh
Department of Computer Science
Korea Advanced Institute of Science and Technology
{cdjin,ysyun,sjyun,yhoh}@bulsai.kaist.ac.kr

요약

본 논문에서는 핵심어 검출의 성능을 향상시키기 위한 새로운 후처리 방법을 제안한다. 일반적으로 핵심어 검출 시스템에 의해 검출된 상위 n개의 후보 단어들의 우도(likelihood)는 비슷한 경우가 많다. 따라서, 한 음성 구간에 대해 음향학적으로 유사한 핵심어들간의 오인식 가능성이 높아진다. 그러나 기존의 핵심어 검출에 사용된 후처리 방법은 음성의 모든 구간에 같은 비중을 두고 우도를 평가하므로 비슷한 음향학적 특징을 가지는 유사한 핵심어들의 비교에 적합하지 못하다. 이를 해결하기 위하여, 본 논문에서는 후보단어들의 구분적인 음향학적 특징 차이에 기반한 가중치를 우도 계산 시에 반영함으로써 보다 변별력을 높이는 알고리즘을 제안한다. 실험 결과, 제안된 방법을 이용하여 유사한 후보단어들간의 변별력을 높일 수 있었고, 인식률이 93%일 때, 우도비검사 방법에 비해 19.6%의 false alarm rate을 감소시킬 수 있었다.

1. 서론

핵심어 검출(word spotting)은 어휘의 제한없이 자연스럽게 발성한 연속음성으로부터 필요로 하는 단어 구간을 검출하여 인식하는 음성인식의 한 방법이며, 사용자에게 발성의 제약을 주지 않고 필요한 몇몇 단어만을 인식함으로써 높은 인식률을 얻을 수 있어 많은 응용분야에서 효과적으로 사용될 수 있다.

초기의 핵심어 검출은 DTW(Dynamic Time Warping)를 사용하였으나, 최근에는 다른 음성인식 분야에서와 마찬가지로 HMM(hidden Markov model)을 이용한 방법이 주류를 이루고 있다. HMM을 이용한 시스템은 일반적으로 인식할 핵심어와 묵음을 포함하는 비핵심어 모델로 구성되며, 입력 음성을 핵심어와 비핵심어의 열을 표현하여 핵심어를 검출하게 된다.

이렇게 검출된 후보 단어는 오인식을 감소시키기 위해 후처리를 통하여 다양한 방법으로 신뢰도를 평가받게 된다. 신뢰도 측정은 핵심어 검출 시스템의 성능을 완벽하게 보장할 수 없고, 오인식으로 인한 피해를 방지하기 위해서이다. 그러나 핵심어 검출 시스템이 검출해 내지 못한 핵심어를 후처리에서 찾아낸다는 것은 매우 어렵기 때문에, 일반적으로 false alarm을 제거하는데 초점을 두고 있다.

본 논문에서는 핵심어 검출의 후처리 과정에서 보다 효과적으로 신뢰도를 측정하고, 검출되지 못했거나 치환(substitution)된 단어를 복원하는 후처리 방법으로 핵심어 검출 시스템의 성능을 높이고자 한다.

본 논문은 총 6장으로 구성되어 있다. 2장에서는 기본적인 핵심어 검출 시스템에 대하여 설명하고, 3장에서는 대표적인 후처리방법인 우도비검사(likelihood ratio test) 방법에 대해 설명하고, 이 방법의 문제점을 지적한다. 다음 4장에서 제안하는 상태(state)별 가중치를 이용한 후처리 방법을 설명 한 후, 5장에서 실험 결과를 제시하고, 6장에서 결론 및 추후 연구사항을 정리한다.

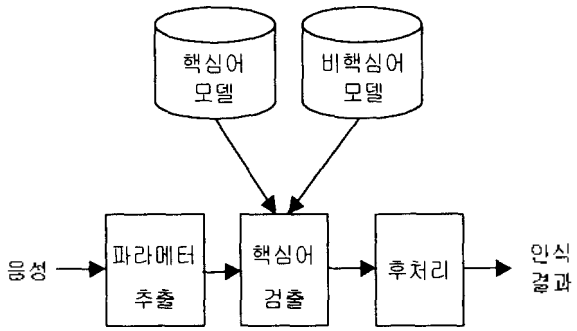

그림 1 핵심어 검출 시스템의 기본 구성

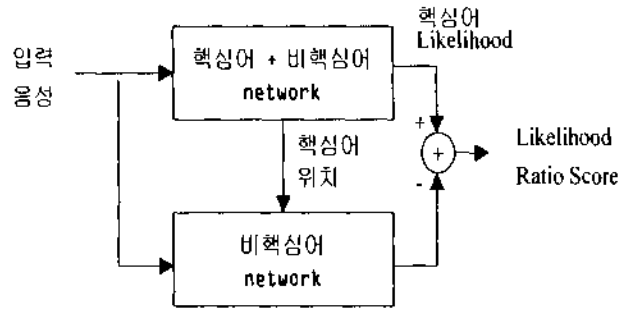

그림 2 우도비검사 방법의 기본 구성도

2. 핵심어 검출 시스템의 기본 구성

핵심어 검출 시스템의 일반적인 구성은 그림 1과 같다. 먼저, 입력음성에서 음성신호의 음향학적 특성을 잘 나타내는 특징 파라메터를 추출한다. 특징 파라메터는 cepstrum이나 인간의 인지 특성을 이용한 mel-cepstrum이 주로 사용된다.

다음, 핵심어 모델과 비핵심어 모델을 이용하여 핵심어들을 검출한다. 핵심어 모델은 단어 전체를 기본 모델로 하는 방법과 음소 모델을 기본 모델로 하고 이들의 연결된 형태로 핵심어를 모델링하는 방법이 있다. 단어 전체를 모델링한 경우 음소간의 음운 변이를 포함시킬 수 있으므로 우수한 인식성능을 얻을 수 있으나 핵심어의 추가나 변경이 용이하지 않다는 단점을 가진다. 반면 음소별로 모델링을 하는 경우, 새로운 핵심어를 추가하거나, 변경할 때 추가적인 훈련자료와 훈련과정이 필요 없다는 장점을 가지지만, 음운 변이를 잘 표현할 수 없기 때문에 인식성능은 떨어진다. 한편, 비핵심어 모델은 핵심어에 해당되지 않는 음성들과 묵음이나 배경잡음을 표현한다. 핵심어 검출 방법은 주로 one-pass DP (dynamic programing)알고리즘이 사용된다. one-pass DP알고리즘은 연결단어인식을 위한 매우 효과적인 방법으로 이용되고 있으며, 프레임 동기(frame-synchronous)로 실시간 시스템 구성에 매우 유리하다. 이 알고리즘은 입력음성에 대해 최적 경로를 찾는다.

마지막으로 후처리 단계에서 검출된 후보단어에 대해 신뢰도 측정을 하게 된다. 후처리의 목적은 오인식의 가능성을 줄이는 데에 있다. 일반적으로 후보단어의 신뢰도가 떨어지는 경우 차라리 인식하지 않는 편이 잘못 인식하는 경우보다 피해를 줄일 수 있다. 신뢰도 측정에는 여러 가지 방법이 있으나, 대부분의 방법에서 임계치에 따라 인식률과 false alarm rate간의 trade-off가 존재한다. 즉, 인식률을 높이려면 false alarm rate의 증가를 감수해야 하고, 반대로 false alarm rate를 감소시키려면 인식률의 저하를 감수해야 한다. 그러므로, 인식률의 저하를 최소한으로 줄이면서, false alarm rate를 감소시키는 후처리 방법이 좋은 것이라 할 수 있다.

3. 기존의 후처리 방법

핵심어 검출 시스템에서 사용되는 후처리 방식으로는 음성 segment를 이용하는 방법[1], 변별적 훈련과정을 사용하는 방법[2], 신경회로망을 이용하는 방법[3], 그리고 우도비검사 방법[4] 등이 있다. 이들중 가장 대표적인 방법은 우도비검사 방법을 들 수 있다.

이 방법은 검출된 후보 단어 구간이 비핵심어 HMM 모델을 통과시킨 우도에 비해 후보 단어 HMM모델을 통과시킨 우도의 차이를 기준으로 후보 단어를 최종적으로 인정할 것이지 여부를 결정한다. 본 논문에서 사용한 우도비검사 방법의 개략도는 그림 2와 같다. 탐색과정에서 후보 단어가 검출되면 후보 단어의 음성구간을 비핵심어로만 구성된 network에 넘겨주고, 후보 단어의 음성구간을 비핵심어 network에 통과시켜 우도를 계산한다. 이 우도와 후보단어 HMM을 통과시켰을 때의 우도와의 차이를 S라고 하면 이것은 식 1과 같다. 여기서 S가 임계치를 넘지 못하면 이 후보 단어는 신뢰할 수 없는 것으로 판단하여 기각시킨다.

$$S = \log P(O_{T_1}^{T_2}|w) - \log P(O_{T_1}^{T_2}|f) \qquad (1)$$

여기서 w와 f는 각각 핵심어 모델과 비핵심어 모델을 나타내며, $O_{T_1}^{T_2}$ 는 T_1에서 T_2까지의 관찰 벡터열을 나타낸다.

그러나 우도비검사 방법은 검출된 후보단어가 핵심어에 가까운지 비핵심어에 가까운지를 살펴보는 것으로, 다른 핵심어로 잘못 인식되는 경우에 대한 고려가 없다. 즉, 비슷한 핵심어들 사이에서 발생하는 치환현상의 경우 우도비검사 방법으로는 효과적으로 제거할 수 없게 된다.

이러한 치환 오류를 제거하기 위하여 anti-keyword

모델이 제안되었다[7]. 이 방법은 후보 단어의 anti-keyword 모델(특정 핵심어 외의 핵심어들로 훈련된 시간 모델)에 대한 우도나 후보단어 이외의 핵심어 모델에 대한 평균 우도를 후보 단어 HMM의 우도와 비교함으로써 거각여부를 가린다. 하지만, 이 방법은 비교 단어들의 부분적인 유사성을 고려하지 않은 채 전체적인 비교를 함으로써, '가자와'와 '가자다와'와 같이 부분적으로 비슷한 음향학적 특성을 나타내는 단어들을 구별하기 어렵다는 문제점을 가진다.

4장에서는 이처럼 음향학적으로 비슷한 단어들간에 일어나는 치환현상을 효과적으로 제거하는 방법을 설명한다.

4. 상태별 가중치에 의한 후처리 방법

검색과정에서 검출된 후보단어는 먼저 우도비검사 방법에 의해 1차적인 검증을 거친 후, 두 번째 후보단어와의 치환여부를 검증하게 된다. 이 때, 두 후보 단어의 구간이 일치하지 않을 수 있으므로, 두 후보단어의 앞뒤로 사상 가깝게 인식되는 비핵심어까지를 비교 구간으로 결정한다.

일반적으로 입력 음성의 특성 HMM에 대한 우도는 식 2로 계산된다.

$$\begin{aligned} g(O;\Lambda) &= \frac{1}{T_2-T_1+1}\log\left[a_{s_{T_1}s_{T_1}}\cdot\prod_{t=T_1}^{T_2}a_{s_t,s_{t+1}}b_{s_t}(o_t)\right] \\ &= \frac{1}{T_2-T_1+1}\left(\log[a_{s_{T_1}s_{T_1}}]\right. \\ &\quad \left.+\sum_{t=T_1}^{T_2}\left(\log[a_{s_t,s_{t+1}}]+\log[b_{s_t}(o_t)]\right)\right) \\ &= \frac{1}{T_2-T_1+1}\left(\log[a_{s_0s_0}]\right. \\ &\quad \left.+\sum_{all\,t}\left(\sum_{t\in S_t}\left(\log[a_{s_{t-1},s_t}]+\log[b_{s_t}(o_t)]\right)\right)\right) \end{aligned} \tag{2}$$

이때, T_1, T_2는 비교 구간이고, $K(o_t)$는 time t에서의 코드vector의 인덱스이며, S_t는 $\{t|s_t \text{ at } t,\ T_1<t<T_2\}$로 정의된다.

그러나, 식 2은 전 구간에 걸쳐 같은 비중을 두고 우도를 계산하게 되므로, 입력음성 전체적인 특성은 잘 반영할 수 있지만, 음향학적으로 비슷한 두 단어를 비교하는 데에는 적합하지 못하다.

음향학적으로 비슷한 두 단어를 비교할 때에는 두 단어의 특성을 잘 나타내며, 차이를 잘 드러내는 특성 구간에 비중을 두고 비교하는 것이 바람직하다.

그러므로, 제안하는 후처리 방법에서는 두 후보 단어 사이의 비교를 효과적으로 수행하기 위하여 각 음성 구간에 대해 가중치를 준다. 이 가중치는 두 후보단어의 비교시에 특성을 잘 나타내는 구간의 비중을 높이고, 그렇지 못한 구간의 비중을 낮추는 역할을 한다. 가중치를 이용하여 식 3에서 수정된 우도를 계산할 수 있다.

$$\begin{aligned} g'(O;\Lambda) &= \frac{1}{T_2-T_1+1}\left(\log[a_{s_0}]\right. \\ &\quad \left.+\sum_{all\,t}\sum_{t\in S_t}\left(\log[a_{s_{t-1},s_t}]+w_t\cdot\log[b_{s_t}(o_t)]\right)\right) \end{aligned} \tag{3}$$

가중치 w_t는 식 4에서와 같이 상태간 거리를 이용하여 계산된다.

$$\begin{aligned} w_t &= \gamma\cdot\frac{d(s_t^{(1)},s_t^{(2)})}{\sum_{t=T_1}^{T_2}d(s_t^{(1)},s_t^{(2)})}\times(T_2-T_1+1) \\ &\quad +\ (1-\gamma) \end{aligned} \tag{4}$$

이때, γ는 가중치의 적용비율이며, γ가 높으면 부분적인 비교를 할 수 있고, 낮으면 전체적인 특성을 반영할 수 있다.

두 후보 단어 사이의 가중치는 두 후보 단어 HMM의 대응되는 상태간 거리에 비례하여 주어지게 되며, HMM의 상태간 거리는 식 5로 구할 수 있다.

$$d(s_t^{(1)},s_t^{(2)}) = \left\{\frac{1}{M}\sum_{k=1}^{M}\left[b_k(t)^{(1)}-b_k(t)^{(2)}\right]^2\right\}^{\frac{1}{2}} \tag{5}$$

이 때, $s_t^{(1)}, s_t^{(2)}$는 각각 두 후보 단어의 시간 t에서의 부분적 상태(state)를 나타내며, M은 코드북 크기, $b_k(t)^{(1)}$, $b_k(t)^{(2)}$는 각각 두 후보 단어의 시간 t에서의 k번째 코드워드의 출력확률을 나타낸다.

이처럼 시간별로 대응되는 상태의 차이를 우도 계산에 반영함으로써 음향학적으로 차이가 적은 부분은 비중을 낮추고, 차이가 큰 부분은 비중을 높여, 부분적인 비교가 가능해 진다.

이 계산과정에서 주의해야 할 점은 가중치를 적용하기 전의 탐색경로가 가중치에 의해 영향을 받지 않아야 한다는 점이다. 만약 가중치에 의해 탐색경로가 바뀔 수 있다면, 높은 가중치를 갖는 상태에서 오래 머물러 탐색경로가 왜곡되는 현상이 발생할 수 있기 때문이다[5]. 본 실험에서는 가중치를 적용하기 전에 탐색한 경로를 그대로 유지하면서 가중된 우도를 계산함으로써, 탐색경로의 왜곡현상을 방지하였다.

위 식 3에서 구해진 두 후보 단어의 가중된 우도를 비교하여 일정 임계치보다 차이가 클 경우에는 높은 우도를 나타내는 단어를 결과로 출력하며, 작을 경우에는 두 단어간의 치환현상을 막기 위해 기각한다.

위와 같은 방법을 사용함으로써, 핵심어사이의 치환을 사전에 방지할 수 있으며, 후보단어를 보다 효과적으로

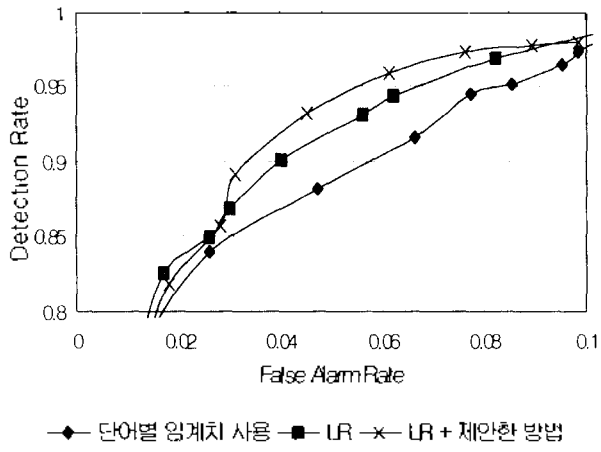

그림 3 제안한 방법의 성능 비교

검증할 수 있다.

5. 실험 및 결과

제안한 방법의 유효성을 알아보기 위해 다수의 화자가 발성한 음성 자료를 이용하여 인식 실험을 하였다.

음성자료는 16kHz로 샘플링되었으며, 15개의 비슷한 핵심어가 평균 2.32개씩 포함된 1440문장에 대하여 실험하였다. 35명의 여성화자와 34명의 남성화자의 발성한 1011문장을 훈련에 사용하였고, 훈련에 사용되지 않은 33명의 여성화자와 10명의 남성화자의 429문장을 대상으로 인식실험을 하였다.

입력값은 256개의 코드워드를 갖는 코드북을 사용하여 양자화하였고, 단어별로 10-18개의 상태를 갖는 left-to-right형 이산분포HMM을 사용하였다. 특성벡터로는 14차 mel-cepstrum, Δmel-cepstrum, energy, Δenergy를 사용하였다. 핵심어 검출에 사용되는 탐색방법은 beam search와 one-pass DP (Dynamic Programing)이다[6].

그림 3은 후보 단어의 우도를 직접 임계치와 비교하는 방법, 우도비검사 방법, 그리고 제안한 방법에 대해 실험한 결과이다. 세로축은 인식률, 가로축은 false alarm rate를 나타내며, 그래프가 위쪽 왼쪽 갈수록 좋은 성능을 나타냄을 의미한다.

그래프에서 보는 바와 같이 제안한 방법이 인식률의 저하가 거의 없이 효과적으로 false alarm을 제거함으로써 기존의 우도비검사 방법보다 높은 성능을 나타내었다. 이것은 가중치를 사용함으로써 유사한 핵심어 사이의 부분적인 비교를 효과적으로 수행할 수 있었기 때문이다.

6. 결론

본 논문에서는 핵심어 검출 시스템에서 보다 효과적으로 후보단어를 검증할 수 있는 후처리 방법을 제안하였다. 음향학적으로 비슷한 핵심어가 평균 2.32개씩 포함된 1440문장을 대상으로 실험한 결과 제안한 방법에 의해 후처리된 핵심어 검출 시스템의 성능이 기존 우도비검사 방법보다 높은 성능을 나타내었다. 이것은 기존의 우도비검사 방법이 후보단어를 비핵심어 모델과만 비교한 것에 비해, 제안한 방법은 다른 후보 단어와도 비교함으로써, 보다 효과적으로 false alarm을 제거하였기 때문이다.

제안된 방법은 후보 단어들간에 차이를 많이 나타내는 구간을 중점적으로 비교함으로써 음향학적으로 비슷한 두 단어를 보다 효과적으로 구분할 수 있다. 아울러, 두 후보간의 비교를 통해 치환을 교정할 수 있는 장점을 지닌다. 또한 이 방법은 고립단어인식시에는 물론이고, 비슷한 음소를 효과적으로 구분함으로써 연속음성인식에도 별다른 수정없이 사용하여 성능을 향상할 수 있다.

제안한 방법에서는 상태간 거리를 이용하여 가중치를 계산하였으나, 후보 단어사이의 특징을 좀더 잘 나타내는 파라메터를 가중치 계산에 사용한다면 성능을 더 개선시킬 수 있을 것이라고 예상된다.

참고문헌

[1] H.Gish and K.Ng, "A segmental speech model with application to word spotting," *Proc. of ICASSP*, Vol. II, pp. 447-450, 1993.

[2] R.A.Sukkar and J.G.Wilpon, "A two pass classifier for utterance rejection in keyword spotting," *Proc. of ICASSP*, Vol. II, pp. 451-454, 1993.

[3] D.P.Morgan, C.L.Scofield and J.E.Adcock, "Multiple neural network topologies applied to keyword spotting," *Proc. of ICASSP*, pp. 313-316, 1991.

[4] R.C. Rose, D.B. Paul, "A hidden markov model based keyword recognition system," *Proc. of ICASSP*, pp. 129-132, 1990.

[5] F.Wolfertstetter and G.Ruske, "Discriminative state-weighting in hidden markov models," *Proc. of ICSLP*, pp. 219-222, 1994.

[6] H. Ney, D. Mergel, A. Noll, and A. Paeseler, "A data-driven search organization for continuous speech recognition," *IEEE Trans. Signal Processing*, vol. 40, No. 2, pp. 272-281, 1992.

[7] M.G.Rahim, C.-H.Lee and B.-H.Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. Speech and Audio Processing*, vol. 5, No. 3, pp. 266-277, 1997.