Keyword spotting에서의 후처리과정에 관한 연구

o
송 화전, 김 재호, 손 경식, 김 형순
부산대학교 전자공학과

A Study on the Postprocessing in Keyword Spotting

o
Hwa Jeon Song, Jae Ho Kim, Kyung Sik Son, Hyung Soon Kim
Dept. of Electronic Engineering, Pusan National University

要 約

Keyword spotting이란 음성인식의 한 분야로서 컴퓨터가 사람의 음성을 입력받아 이 음성에 미리 정해진 특정단어(keyword) 또는 특수개의 단어들 중 어느 것이 포함되어 있는지의 여부를 찾아내고 이 단어를 식별해 내는 작업을 의미한다. 이러한 keyword spotting 시스템의 인식 오류들을 감소시키는 방법의 하나로 keyword spotting 시스템에 후처리 과정을 통으로써 잘못 검출된 keyword들을 제거시키는 방법이 사용될 수 있다.

본 논문에서는 keyword로 검출된 영역에 대한 keyword 모델의 likelihood와 그 영역에 대한 filler 모델의 likelihood의 ratio와 second best keyword의 likelihood 그리고, 끝점존재 영역의 구간 길이등 여러 가지 정보를 이용한 후처리과정을 검토하고 인식실험을 통해 이들의 성능을 비교하였다. 6개의 부서명을 keyword로 하는 불특정화자 keyword spotting 실험을 수행한 결과 baseline 시스템의 경우 고립단어 및 문장형태의 음성에 대해 95.0%의 keyword 인식률을 얻었으며, 본 논문에서 검토된 네 가지 후처리 방법에 의해 keyword rejection ratio를 0%에서 5%까지 변화시켜 나갈 경우 최저 95.3%에서 최고 97.1%까지 keyword 인식률이 향상된 결과를 얻었다. 특히, 성능과 계산량을 종합적으로 고려할 때 끝점 존재 영역의 구간 길이 정보를 이용한 방법이 가장 우수하였다.

Ⅰ. 서 론

음성인식은 크게 고립단어인식과 연속음성인식, 그리고 keyword spotting의 세 가지 부류로 나눌 수 있다. 고립단어인식은 구현 면에서 용이하고 비교적 높은 인식률을 얻을 수 있지만, 사용자가 또박또박 끊어 발음해야 하는 부자연스러움을 감수해야 한다. 연속음성인식은 사용자에게는 자연스럽지만 매우 제한된 응용분야를 제외하고는 아직까지는 인식성능이 기대수준에 크게 뒤떨어지므로 실제환경에 사용하기가 어렵다. 이에 반하여 keyword spotting은 미리 특정단어(keyword)를 지정해 놓은 후 사람들이 자연스럽게 말한 연속음성으로부터 이들 keyword들을 추출해 내는 것으로서, 고립단어인식에서의 사용자의 불편함과 연속음성인식에서의 성능 저조의 문제점을 모두 해결할 수 있는 방식이다. 따라서, keyword spotting은 문장 내에서 핵심 주제어만 검출해 내면 의미가 통할 수 있는 많은 응용분야에 효과적으로 사용될 수 있으며, 사용자의 편리함이 강조되는 추세와 더불어 그 역할의 중요성이 점차 증대되고 있다.

인식대상 단어들에 대한 인식률 또는 오인식률을 성능평가 기준으로 삼는 고립단어인식의 경우와는 달리 keyword spotting에서는 keyword의 다른 keyword로의 오인식, keyword를 검출하지 못한 경우, 그리고 keyword가 아닌 부분을 keyword로 잘못 검출하는 경우(false alarm)등의 세가지 오류가 발생할 수 있다. 이러한 오류들을 감소시키는 방법으로서 likelihood ratio scoring 방법을 비롯한 몇 가지의 후처리 방법이 개발되어 왔으며[2][4][5], 이들은 keyword spotting 시스템이 검출해 내지 못한 keyword를 재검출하기보다 이미 구해진 keyword 후보들로 부터 false alarm을 효율적으로 제거하는데 주안점을 두고 있다. 따라서, 이들 후처리 방법들은 검출된 keyword들의 신뢰도를 적절한 방법으로 평가하여 신뢰도가 떨어진다고 판단되는 후보들을 제외시킴으로써, 잘못 검출된 keyword에 의해 야기될 수 있는 혼란을 미리 방지하게 된다.

본 논문에서는 부서 전화번호 안내 및 교환업무라는 task domain을 설정하여 불특정 화자가 아무런 부법적 제한없이 발음한 연속음성으로부터 부서명(keyword)를 찾아 내어 사용자에게 부서에 대한 정보를 알려주는 keyword spotting 시스템을 baseline 시스템으로 구축하고, 이 시스템의 성능을 향상시키기 위한 방법으로서 keyword 검출시 나타나는 여러가지 정보를 이용한 몇 가지의 후처리 방법을 검토하여 인식실험을 통해 그 성능을 비교하였다.

본 논문의 구성은 다음과 같다. 2장에서는 baseline keyword spotting 시스템에 대해 서술하고, 3장에서는 여러가지 정보를 이용한 후처리 과정에 대해 기술한다. 그리고 4장에서는 후처리과정을 수행한 실험결과를 기술하고, 마지막으로 5장에서 결론을 맺는다.

II. Keyword Spotting Baseline 시스템

A. keyword spotting 시스템의 기본 구조

본 논문에서 구현된 keyword spotting 시스템의 구성은 그림 1과 같다. 음성 전처리 과정에서는 음성신호의 특징 파라메타를 추출하고, keyword 검출과정에서는 추출된 파라메터를 미리 만들어진 keyword 모델 및 keyword가 아닌 음성 부분, 즉, non-keyword의 모델, 그리고 묵음구간을 나타내는 silence 모델과 비교하여 keyword 후보를 찾아내다. 그리고, 마지막으로 후처리 과정에서 검출된 keyword의 신뢰도를 평가하여 잘못 검출된 keyword들을 제거시킴으로써 시스템의 성능을 향상시킨다.

B. 음성신호 전처리 과정

음성신호는 16kHz로 샘플링하여 전달함수가 $1-0.97z^{-1}$인 1차 필터로 preemphasis를 한 다음, 길이가 20msec이고 10msec씩 중첩되는 frame 단위로 나누어 Hamming window를 씌운다. 매 frame마다 자기상관 방법에 의한 LPC분석을 한 다음 이로부터 구한 12개의 LPC cepstrum들과 음성신호의 시간축 상에서의 정보를 보존하기 위해 선형 회기분석 방법을 이용하여 구한 12개의 cepstrum derivative(또는 delta cepstrum)를 음성특징 파라메터 벡터로 사용한다.

C. HMM 기본구조

HMM의 기본적인 형태는 1차 left-to-right Markov 모델을 토대로 하고 있으며, 상태천이 행렬 $A = \{ a_{ij} \}$는 다음과 같은 제약조건을 갖도록 하였다.

$$a_{ij} = 0 \qquad j < i,\ j > i+2 \tag{1}$$

또한 관찰확률행렬 $B = \{ b_j(x) \}$는

$$b_j(x) = \sum_{m=1}^{M} c_{mj} N(x, \mu_{mj}, U_{mj}) \tag{2}$$

로 주어지며, 이 때 $N(\ )$은 Gaussian 확률밀도함수이고 x는 해당 frame의 파라메터 벡터이다. 그리고, c_{mj}는 상태 j에서의 m번째 mixture에 해당하는 가중치이며, u_{mj}와 U_{mj}는 각각 상태 j에서의 m번째 mixture의 평균 벡터 및 covariance이다. 여기서 covariance matrix는 주대각 항들만을 사용하였으며, 각 상태 당 mixture의 갯수(M)는 9개로 정하였다. Keyword HMM들의 상태 갯수는 음절 수를 고려하여 7개에서 10개 사이에서 가변적으로 결정하였으며, non-keyword 및 silence HMM의 상태 갯수는 10개로 정하였다.

본 논문에서는 bootstrapping 방법을 사용하여 keyword와 filler HMM들이 자동적으로 생성되도록 하였으며 그 과정은 다음과 같다. 먼저 고립단어 형태의 keyword 음성 데이타베이스로부터 1차적인 keyword HMM을 훈련시킨다. 또한 이 데이타베이스에서 자동 음성/비음성 검출과정을 통해 구해진 비음성 구간으로부터 1차적인 silence HMM을 훈련시킨다. Non-keyword HMM은 임차적으로 random 데이타로부터 구한다. 이와 같이 1차적인 keyword, silence 및 non-keyword모델이 구성되면, 이들을 이용하여 문장형태의 keyword spotting 훈련용 음성 데이타베이스로부터 Viterbi decoding에 의한 자동분할을 행한다. 이와 같은 음성 데이타의 자동 분할과 그 결과에 따른 keyword, non-keyword 및 silence HMM들의 재훈련과정을 수렴상태에 이를 때까지 반복적으로 수행시킨다[1].

D. 후처리 과정

인식대상 단어들에 대한 인식률 또는 오인식률을 성능평가 기준으로 삼는 고립단어인식의 경우와는 달리 keyword spotting에서는 keyword의 다른 keyword로의 오인식, keyword를 검출하지 못한 경우, 그리고 keyword가 아닌 부분을 keyword로 잘못 검출하는 경우(false alarm)등의 세 가지 오류가 발생할 수 있다.

후처리 과정은 이러한 오류들을 감소시켜 그 성능을 보다 향상시키기 위한 것으로서, HMM에 의한 keyword 검출과정에서 검출해 내지 못한 keyword를 후처리 과정에서 찾아낸다는 것은 많은 문제가 따르기 때문에, 후처리 과정은 일반적으로 이미 구해진 keyword 후보들의 신뢰도를 적절한 방법으로 평가하여 잘못 검출된 keyword를 효율적으로 제거하는데 주안점을 두고 있다. 후처리 과정의 구체적인 방법에 대해서는 다음 장에서 보다 더 상세히 설명하겠다.

III. 후처리 과정의 방법

지금까지 keyword spotting 시스템의 성능개선을 위해 시도되어온 후처리 방법으로는 segmental 모델을 이용하는 방법[4], generalized probabilistic descent(GPD) 변별훈련과정에 의한 방법[5], 그리고 likelihood ratio scoring 방법[2]등이 있으나, 이들 중 첫번째 및 두번째 방법은 후처리를 위해 별도의 새로운 파라메터들을 구해야 하는 등의 불편함이 따른다. 본 논문에서는 liklihood ratio scoring 방법을 비롯하여 keyword spotting baseline 시스템에서 나올 수 있는 여러가지 정보들을 사용한 후처리 방법들을 검토하였다.

A. Likelihood Ratio Scoring 방법[2]

이 방법은 그림 2에 나타낸 바와 같이 두 가지의 HMM network을 병렬로 사용한다. 그 중 첫번째는 keyword 및 filler network이고, 두번째는 keyword 모델없이 filler 모델만으로 구성된 network이다. 여기

서 filler 모델이란 non-keyword 및 silence 모델을 통칭한 표현이다. keyword 및 filler network는 그림 1의 keyword 검출부에 해당하는 것으로서 입력문장으로부터 keyword 후보 및 그 위치를 검출해 내고, filler 모델만으로 구성된 network는 keyword 후보위치에 해당하는 filler string을 구할 수 있게 된다. 이 때, keyword의 log likelihood로부터 이 구간에 해당하는 filler 모델의 log likelihood를 뺀 값을 likelihood ratio score라고 한다. 이 score를 적절한 경계치와 비교함으로서 keyword 검출여부를 최종적으로 결정한다.

B. Likelihood Ratio Scoring 방법의 변형 - 1

이 방법은 A 방법과 거의 유사하지만 한가지 다른 것은 filler모델의 log likelihood를 계산하는데 있어서 filler 모델만으로 구성된 network에 문장 전체를 통과시키지 않고 단지 keyword로 검출된 영역만을 통과시켜 filler log likelihood를 계산한다. 이렇게 함으로서 A 방법에 비교하여 많은 계산량을 줄일 수 있다.

C. Likelihood Ratio Scoring 방법의 변형 - 2

Keyword spotting에서 검출된 keyword를 best keyword라 하고 이를 제외한 keyword들만 가지고 다시 keyword spotting을 수행했을 때 검출된 keyword를 2nd best keyword라 할 때, 이 방법에서는 best keyword log likihood와 2nd best keyword log likelihood의 차이를 이용한다. 이 방법은 A 방법 및 B 방법에 비해 더 많은 계산량이 소요된다.

D. 끝점 존재 영역의 길이에 의한 방법

Keyword 검출과정에서 Viterbi decoding에 의해 문장을 분할할 때 매 frame마다 가장 높은 likelihood값을 가지는 모델이 어느 것인가를 저장해 놓는다. 어려한 데이타는 음성을 분할한 후 검출된 keyword가 다른 모델들에 비해 어느 영역에서 가장 높은 likelihood값을 가지는지에 대한 정보를 제공한다. 검출된 keyword의 끝점 부근에서 keyword 모델이 다른 모델들 보다 높은 likelihood를 나타내는 영역을 끝점 존재 영역이라 부르기로 한다[3]. 이 끝점 존재 영역의 길이를 조사해 보면 keyword를 잘못 검출할 경우가 올바로 검출한 경우에 비해 상대적으로 짧은 길이를 갖게 된다. 따라서, 끝점 존재 영역의 길이를 적절한 경계치와 비교하여 keyword 검출여부를 결정하게 되며, 이 방법은 앞의 세가지 방법에 비해 가장 적은 계산량을 필요로 한다.

IV. 실험 및 결과

본 논문에서는 부서 전화번호 안내 및 교환업무를 keyword spotting 시스템의 대상으로 선정하였다. 이에 따라 6개의 부서명을 keyword로 선정하여 남성화자 50명이 각 keyword가 1개씩 들어있는 문장을 1번씩 발음하게 하고, 또한, 고립단어 형태로도 1번씩 발음하게 한 것을 음성 데이타베이스로 사용하였다[3].

6개의 keyword들을 대상으로 화자독립 keyword spotting 실험을 수행하기 위해 먼저 35명이 발음한 고립단어 형태와 문장형태의 음성데이타를 훈련에 참여시켰고, 훈련에 참여하지 않은 15명이 발음한 고립단어 형태와 문장형태의 음성데이타로 인식실험을 하였다. 후처리과정을 제외한 baseline keyword spotting 시스템에 대한 실험결과, 고립단어 형태와 문장형태의 음성에 대해 각각 97.8% 및 92.0%의 keyword 인식률을 얻었으며 전체 keyword 인식률은 95.0% 였다.

그리고, III장에서 서술한 네 가지의 후처리 방법을 도입하였을 때의 실험 결과가 그림 3에 나타나 있다. 본 논문에서는 각각의 방법에 대해 경계치를 점차 상향조정시킴으로써 인식대상 keyword들이 기각(rejection)되는 비율이 0%에서 5%로 변화되는 과정에서의 후처리 방법의 성능을 비교하였다.

그림 3(c)에서 보는 바와 같이 방법 C를 제외하고는 keyword 기각률이 0%에서 5%로 증가됨에 따라 keyword 인식률이 상당히 향상되는 것을 알 수 있다. 그리고, 방법 B와 방법 D는 방법 A에 비해 계산량이 적게 소요되면서도 동등하거나 오히려 약간 우수한 성능을 나타내고 있으며, 그림 3(b)에 나타난 바와 같이 특히 문장형태의 음성 데이타로부터 잘못 검출된 keyword를 효과적으로 제거할 수 있음을 보여주고 있다. 그 중에서도 방법 D는 최소한의 계산량으로 가장 우수한 성능을 나타내고 있으며, 따라서 끝점 존재 영역의 길이에 의한 후처리 방법이 계산량 및 성능을 종합적으로 고려할 때 가장 우수한 방법임을 시사한다.

V. 결 론

본 논문에서는 부서 전화번호 안내 및 교환업무라는 task domain을 설정하여 불특정 화자가 아무런 문법적 제한없이 말한 연속음성으로부터 부서명(keyword)을 찾아 내어 사용자에게 부서에 대한 정보를 알려주는 keyword spotting 시스템을 baseline 시스템으로 구축하였고, keyword spotting에서 발생한 false alarm을 효율적으로 제거하여 시스템의 성능을 보다 더 향상시키기 위해 keyword 검출시 나타나는 여러가지 정보를 이용한 몇 가지의 후처리 방법을 검토하고, 각각의 방법들에 대해 그 성능비교를 하였다. 실험결과 후처리 방법을 도입하지 않았을 경우, 고립단어 형태와 문장 형태의 음성에 대해 각각 97.8% 및 92.2의 keyword 인식률을 얻었으며 전체 keyword 인식률은 95.0% 였다. 그리고, keyword로 검출된 영역에 대한 keyword 모델의 likelihood와 그 영역에 대한 filler 모델의 likelihood ratio와 second best keyword의 likelihood 그리고, 끝점존재 영역의 구간 길이 등의 다양한 정보를 이용하여 후처리 과정을 수행한 결과, keyword rejection ratio를 0%에서 5%까지 변화시켜 나갈 경우 95.3%에서

97.1%까지 keyword 인식률이 향상된 결과를 얻었다. 특히, 성능과 계산량을 고려할 때 끝점 존재 영역의 구간 길이 정보를 이용한 방법이 가장 우수한 방법으로 나타났다. 현재 이들 후처리 방법들을 종합적으로 사용하여 keyword spotting 시스템의 성능을 보다 더 향상시키고자 하는 연구가 진행중에 있다.

참 고 문 헌

[1] J. G. Wilpon, L. R. Rabiner, C.H. Lee and E.R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," IEEE Trans. Acoust., Speech, Signal Processing, vol.38, no.11, pp1870-1878, Nov. 1990.

[2] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in Proc. IEEE ICASSP, 1990, pp. 129-132.

[3] 송 화전, 강 재호, 손 경식, 강 형순, "자동분할에서의 단어경계 보상을 통한 keyword spotting 시스템의 성능 개선," 1994년도 제 7 회 신호처리합동학술대회 논문집, pp.304-307.

[4] H. Gish and K. Ng, "A segmental speech model with applications to word spotting," in Proc. IEEE ICASSP, 1993, pp. II-447-450

[5] Rafid A.Sukkar and Jay G. Wilpon, "A Two Pass Classifier For Utterance Rejection in Keyword Spotting," in Proc. IEEE ICASSP, 1993, pp. II-451-454.

* 본 논문은 한국전자통신연구소 자동통역연구실의 위탁연구 과제 연구 결과의 일부임.

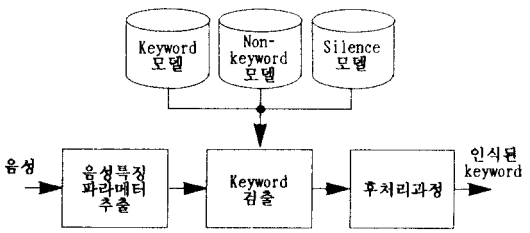


그림 1. Keyword Spotting System의 구성도.

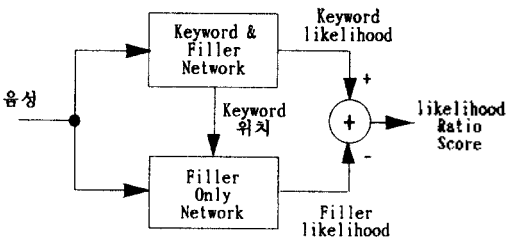
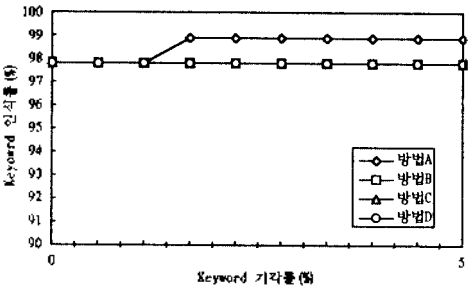
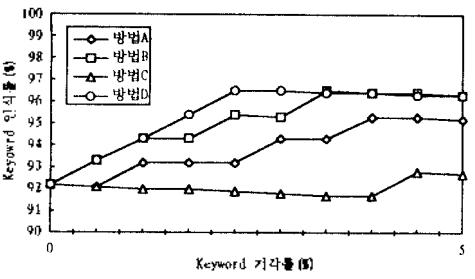


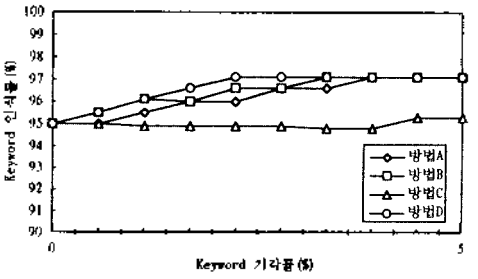
그림 2. Likelihood Ratio Scoring을 이용한 후처리 방법



(a) 고립단어 음성 데이타에 대한 후처리과정의 결과



(b) 문장형태 음성 데이타에 대한 후처리과정의 결과



(c) 전체 음성 데이타에 대한 후처리과정의 결과

그림 3. 후처리 방법들의 성능비교