음소 HMM을 이용한 Keyword Spotting 시스템에서의 Non-keyword 모델에 관한 연구

이 활림°, 김 재호, 손 경식, 김 유신, 김 형순
부산대학교 전자공학과

A Study on the Non-keyword Models in the Keyword Spotting System using the Phone-Based Hidden Markov Models

Hwal Rim Lee, Jae Ho Kim, Kyung Sik Son, Yoo Shin Kim, Hyung Soon Kim
Dept. of Electronics Eng., Pusan National Univ.

요 약

Keyword spotting이란 음성인식의 한 분야로서 입력된 음성에서 미리 정해진 특정단어(keyword) 또는 복수 개의 단어들 중 어느 것이 포함되어 있는지의 여부를 찾아내고 이 단어를 식별해 내는 작업을 의미한다. 음소모델을 이용하여 keyword spotting 시스템을 구성할 경우 새로운 keyword의 추가 또는 변경이 필요할 때 단순히 그 발음사전에 따라 음소모델들을 연결시킴으로써 keyword 모델을 구성할 수 있으므로 단어모델에 의한 방법에 비해 장점이 있다. 본 논문에서는 triphone을 기본단위로 하는 HMM에 의해 keyword 모델을 구성하고, non-keyword 모델 및 silence 모델을 함께 사용하는 keyword spotting 시스템을 구성하였다. 이러한 시스템에서 non-keyword 모델은 keyword와 keyword가 아닌 음성을 구분 지어 주는 역할을 하므로 인식성능의 향상을 위해서는 적절한 non-keyword 모델의 선택이 필요하다. 본 논문에서는 10개의 state를 갖는 단일모델, 조음방법에 의해 음소들을 clustering한 모델, 그리고 통계적 방법에 의해 음소들을 clustering한 모델들을 각각 non-keyword 모델로 사용하여 그 성능을 비교하였다. 6개의 keyword를 대상으로 한 화자독립 keyword spotting 실험결과, 통계적 방법에 의해 음소들을 6 또는 7개의 그룹으로 clustering한 방법이 가장 우수한 인식성능을 나타냈다.

Ⅰ. 서 론

음성인식의 궁극적인 목표는 잡음이 있는 실제적인 환경에서 불특정 화자가 자연스럽게 발음한 대용량 어휘의 연속음성을 실시간에 인식 및 이해하는 것이라고 할 때, 선진 각국의 오랜 연구 노력에도 불구하고 아직까지 이 목표는 달성되지 못하고 있다. 따라서, 음성인식에 관한 대부분의 연구들이 이러한 목표 중 많은 부분에 제약조건을 둔 상태에서 진행되고 있다. 음성인식은 입력 음성의 형태에 따라 고립단어인식과 연속음성인식의 두 가지로 크게 나눌 수 있다. 그 중 고립단어인식은 단어 경계가 분명하므로 인식성능이 우수하지만 사용자가 발음상의 부자연스러움을 감수해야 한다. 이에 반하여 연속음성인식은 문장 형태로 자연스럽게 발음한 음성을 인식하는 것으로서, 사용자의 입장에서는 바람직하지만 아직까지 기술 수준의 한계로 인하여 매우 제한된 어휘와 문법구조를 갖는 경우를 제외하고는 인식성능이 크게 뒤떨어지는 형편이다.

Keyword spotting이란 이들 두 가지 방식들 이외의 제 3의 방식으로서, 어휘에 제한 없이 자연스럽게 발음한 연속음성으로부터 미리 정해진 특정단어(keyword)들을 검출해 내는 기술이다. 따라서 keyword spotting은 고립단어인식에서의 사용자의 불편함과 연속음성인식에서의 성능 저조의 문제를 모두 해결할 수 있으며, 입력된 연속음성으로부터 핵심주제어만 검출해 내면 의미가 통할 수 있는 많은 응용분야, 예를 들면, 전화교환 및 안내 서비스나 각종 정보검색 서비스 등에 효과적으로 활용될 수 있다[1].

HMM을 이용한 keyword spotting 방식은 일반적으로 keyword 모델과 filler 모델들을 사용하는 연결단어인식 알고리즘을 기반으로 하고 있으며[2]-[4], 입력 음성을 keyword 및 filler 들의 시간순서열로 표현하는 과정에서 keyword를 검출하게 된다. 여기서 filler 모델이란 keyword에 해당되지 않는 음성구간, 즉, non-keyword 구간들과 음성이 아닌 배경잡음 구간들을 표현하는데 사용된다.

Keyword 모델들을 구성하는 방법은 크게 단어를 기본단위로 하는 HMM과 음소를 기본단위로 하는 HMM으로 나눌 수 있다. 단어 HMM에 의한 keyword spotting 방식은 각각의 단어 모델들을 훈련시키기 위한 많은 양의 음성 데이타베이스를 필요로 하므로 keyword의 추가 및 변경이 어렵다는 문제점을 지닌다. 이에 반하여 음소를 기본단위로 하는 HMM에서는 모든 단어들이 미리 정해진 음소들의 network 형태로 표현되며, 잘 훈련된 음소 HMM들이 구축되면 새로운 keyword라도 단지 음소 HMM을 연

결시켜 만들면 된다. 따라서 음소 HMM을 이용해 구축된 keyword spotting 방식은 keyword의 추가 및 변경이 용이하게 되어 응용면에서 매우 효과적이다.

Keyword 및 filler 모델에 의한 keyword spotting 방식에서는 filler 모델이 keyword 음성부분을 침해시키지 않으면서 non-keyword 음성 부분 및 배경 잡음 부분을 얼마만큼 효과적으로 표현해 줄 수 있는가에 keyword spotting 방식의 성능이 크게 좌우된다. 지금까지 연구된 filler 모델 구현 방법으로는 non-keyword 각각을 구체적으로 모델링하는 방법[5][4]과 keyword가 아닌 부분 전체를 모델링하는 방법[2]-[4]이 있다. 이들 모델들은 다시 단어 모델 및 subword unit에 의한 모델로 나누어 지며 subword unit을 사용할 경우도 다시 몇 가지 방법들로 구분될 수 있다[3].

본 논문에서는 음소 HMM에 의한 keyword spotting 시스템을 구현하고 인식성능 향상을 위해 몇 가지 non-keyword 모델 구성방법을 검토하였다. 화자독립 keyword spotting 실험 결과 음소를 통계적으로 clustering하는 방법이 가장 우수한 성능을 나타내었다. 본 논문의 구성은 다음과 같다. 서론에 이어 2절에서 baseline keyword spotting system에 대한 설명을 하고 3절에서 non-keyword 모델의 개선에 대해 설명한다. 그리고 4절에서 실험 내용과 결과를 설명하고 마지막으로 5절에서 결론을 맺는다.

II. Keyword Spotting Baseline 시스템의 구성

본 논문 keyword spotting 시스템의 전체적인 구성도가 그림 1에 나타나 있다. 먼저 입력 음성이 들어오면 전처리 과정에서 특징 파라메터들을 추출한다. 음성 데이타가 훈련용일 경우 이로부터 keyword 모델 및 filler 모델(non-keyword 모델과 silence 모델)을 구성하고 이들 모델과 문법적 제약 정보를 이용하여 전체 HMM network을 구성한다. 인식시에는 전처리 과정을 거친 후 HMM network 상에서 Viterbi decoding 과정을 통해 keyword spotting을 수행한다.

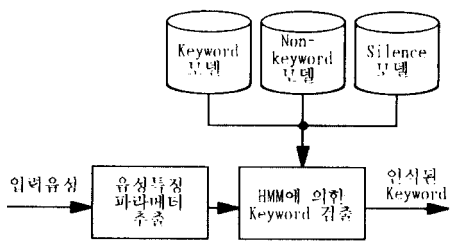

그림 1. Keyword spotting 시스템의 구성도

2.1. 음성신호 전처리 과정

음성신호는 16 kHz로 샘플링하여 전달함수가 $1-0.97z^{-1}$ 인 1차 디지털 필터로 preemphasis를 한다. 그리고 길이가 20 msec이고 10 msec씩 중첩되는 frame 단위로 나눈 다음, 각각의 frame에 Hamming window를 씌운다. 매 frame마다 자기상관 방법에 의한 LPC 분석을 한 다음 이로부터 12개의 LPC cepstrum들을 구한다. 그리고 음성신호의 시간축 상에서의 정보를 보존하기 위해 선형 회기분석 방법에 의한 12개의 delta cepstrum을 구하여 총 24개의 파라메터로 음성특징벡터를 구성하였다.

2.2. 음소 HMM 모델

HMM을 이용한 keyword spotting 시스템의 구현에 있어서 고려해야 할 사항으로는 기본 모델링 단위 정의, HMM topology 정의, 출력확률 분포의 선정, 그리고 각 모델들의 훈련과정 등을 들 수 있다. 본 논문에서는 한국어에 대해 총 46개의 문맥독립형(context-independent) 유사음소를 정의하고 이를 기준으로 한 triphone을 인식단위로 사용하였다. 그리고 본 논문에서 사용한 각 음소 HMM 모델의 topology는 그림 2와 같다. 그림에서 보는 바와 같이 이 모델은 3개의 state와 8개의 transition으로 구성되며 관찰벡터가 각 transition에서 출력된다. 또한 관찰확률분포는 그림과 같이 B, M, E의 3가지 분포로 tying시켰다. 또한 각 transition에서 출력되는 관찰벡터의 확률분포는 state당 복수 개의 mixture를 가질 수 있는 연속확률분포 HMM으로 구현하였으며, state당 mixture의 갯수는 인식실험 결과에 따라 최적화시키도록 하였다.

훈련에 사용한 데이타는 남성화자 35명이 22개 부서명을 각 1회씩 발성한 것을 사용하였다. 각각의 음소모델을 위한 초기모델은 음성학적으로 균형잡힌(phonetically balanced) 445단어에 대한 음성 데이타(22명의 남성화자)로부터 구성하였으며, 부서명 음성 데이타를 이용하여 keyword에 해당하는 triphone 모델을 구성하였다.

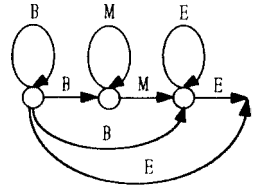

그림 2. 음소 HMM 구조

2.3. Keyword Spotting을 위한 전체 HMM 구조

본 논문에서 구성한 keyword spotting 시스템은 keyword 모델, non-keyword 모델, silence 모델들로 구성된다. 먼저

keyword를 만들기 위한 음소모델은 좌우의 음소까지 고려한 triphone을 사용하였고 이 음소모델들을 연결시켜 keyword에 대한 단어모델을 만들어 keyword spotting 시스템에 사용하였다. 그리고 baseline 시스템에서의 non-keyword 모델과 silence 모델은 각각 keyword가 아닌 음성부분과 묵음 구간을 10 state의 단일 HMM 모델로 구성하였는데, 음소 topology에 맞추기 위해 transition에서 관찰벡터가 나오도록 하였다. Keyword spotting을 위해서는 이들 모델들을 이용하여 전체 HMM network을 구성해야 한다. 일반적으로 한 문장 내에는 임의의 keyword 갯수가 올 수 있으므로 null grammar 형태가 가능하지만, 본 논문에서는 입력음성에 1개의 keyword가 존재한다는 가정하에 그림 3과 같은 문법구조를 가지는 network을 구성하였다[2]. 그림에서 KW, NKW, Sil는 각각 keyword, non-keyword 그리고 silence를 나타낸다.

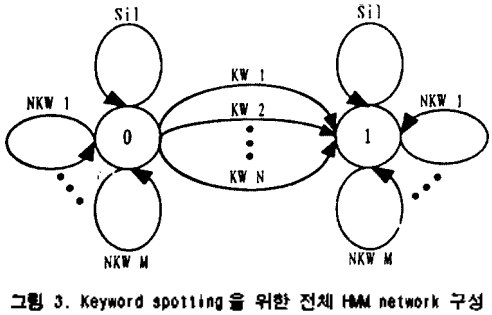

그림 3. Keyword spotting을 위한 전체 HMM network 구성

III. Non-keyword 모델 개선 방법

Filler model을 이용하는 keyword spotting 방식에 있어서 non-keyword 모델은 keyword와 keyword가 아닌 음성을 구분 지어 주는 역할을 한다. 따라서 인식성능의 향상을 위해서는 적절한 non-keyword 모델의 선택이 필요한데, keyword 부분을 잡식하지 않으면서 keyword가 아닌 부분을 어느 정도 표현해 주는 모델이어야 한다. 먼저 baseline keyword spotting 시스템에서는 non-keyword를 10개의 state를 갖는 단일 HMM으로 모델링하였다. 이와 비교하여 문맥 독립형 음소, 즉, monophone들을 clustering하여 non-keyword 모델로 사용하였는데, 모든 monophone을 그대로 사용할 경우 keyword 부분을 잡식할 우려가 있고, 또 모든 음소를 묶어서 하나의 모델로 사용할 경우 keyword가 아닌 부분을 제대로 표현해 주지 못하게 되므로 적절한 grouping이 필요하게 된다.

유사한 특성의 음소를 grouping하는 방법으로 음성학적 지식을 이용하는 방법과 통계적인 방법이 사용될 수 있다. 먼저 음성학적 지식을 이용할 경우 음소들은 조음방법(manner of articulation)에 의하거나 조음위치(place of articulation)에 의해서 분류될 수 있는데, 본 논문에서는 acoustic 특성이 반영된 조음방법에 의해서 음소들을 몇 그룹으로 나누어 non-keyword 모델로 사용하는 방법을 검토하였다. 본 논문에서는 구체적으로 monophone들을 조음방법에 의해서 5개의 그룹으로 나누었는데, 이들은 파열음 그룹, 마찰음 그룹, 파찰음 그룹, 비음, 유음 및 유성자음 그룹, 그리고 모음 그룹이다. 그리고 통계적 방법에 의한 monophone clustering 방법을 검토하였는데, 이 방법은 각각의 모델들 사이의 distance를 정의한 다음, 이 distance가 작은 음소끼리 묶는 방법이다. 본 논문에서는 먼저 음소적으로 균형이 잡힌 445단어의 음성데이타로부터 mixture 1개를 가지는 46개의 유사음소단위 monophone 모델을 만든 후, 각 음소의 확률분포로부터 음소들간의 distance를 구하였다. 그리고 modified k-means(MKM)알고리즘[6]에 의해서 monophone들을 몇 개의 그룹으로 나누었다.

본 논문에서 사용한 단일 mixture를 가지는 음소모델들 사이의 distance measure는 다음과 같이 정의하였다.

$$D(p_i, p_j) = \sum_{d=1}^{N} D_d(p_i, p_j) \qquad (1)$$

여기서 P_i, P_j는 각각 i와 j번째 음소를 나타내고, N은 음소모델의 distribution 수를 나타내며, $D_d(p_i, p_j)$는 두 음소의 각 distribution간의 distance로서 다음 식과 같이 주어진다.

$$D_d(p_i, p_j) = \frac{1}{V}\sum_{k=1}^{V} \frac{(\mu_{idk} - \mu_{jdk})^2}{\sqrt{\sigma_{idk}^2\,\sigma_{jdk}^2}} \qquad (2)$$

이 때, V는 음성특징벡터의 dimension이고 μ_{idk}, μ_{jdk}, σ_{idk}^2 및 σ_{jdk}^2은 각각 i번째 및 j번째 음소의 d번째 distribution에서의 k번째 음성특징 파라메터의 평균 및 분산을 의미한다. 식에서 알 수 있는 바와 같이 각각의 모델 내에서의 transition 확률에 의한 영향은 무시하였다.

IV. 실험 및 결과

음소 모델에 의한 keyword spotting 시스템의 성능을 평가하기 위해 6개의 keyword를 선정하여 keyword spotting 실험을 수행하였다. 본 논문에서 사용한 keyword들은 자동전화교환 서비스를 위한 부서명으로서 총무과, 자산관리과, 회계과, 내자과 설비과 그리고 서울사무소이다. 50명 남성화자의 음성 중에서 36명의 음성을 훈련용으로 그리고 나머지 15명의 음성을 인식실험용으로 사용하였으며, 화자 당 각 부서명에 대해서 1개씩의 고립단어와 문장형태의 음성데이타가 있다. 또한 keyword 인

식과정에서는 그림 3과 같은 HMM network에 대해 Frame-Synchronous Network Search(FSNS) 알고리즘을 사용하였다[7]

음소모델에 의한 baseline 시스템의 실험 결과가 표 1에 나타나 있다. 참고로 단어모델에 의한 keyword spotting 성능도 함께 나타내었는데[8], 문장형태의 데이타에 대해서는 음소모델을 사용하는 경우가 단어모델을 사용하는 경우에 비해 인식 성능이 상대적으로 저조하였다. Baseline keyword spotting 시스템(10개의 state를 갖는 단일 non-keyword모델 구성방법)와 성능을 향상시키기 위한 방법으로서 non-keyword 모델을 구성하는 두 가지 방법, 즉, 조음방법에 의한 monophone clustering 방법과 통계적 방법에 의한 monophone clustering 방법들을 keyword spotting 실험을 통해 비교해 보았다. 통계적 방법에 의한 monophone clustering 방법의 경우 cluster 갯수를 변화시키면서 인식실험을 하였는데, 표 2는 cluster 수에 따른 인식률을 나타내며 cluster 수가 6 또는 7일 때 고립단어 및 문장형태의 음성데이타에 대해 각각 100% 및 93.3%로 가장 좋은 인식률을 보였다. 그러나 조음방법에 의한 monophone clustering 방법의 경우는 음성학적인 지식이 필요하고 특히 cluster 갯수를 바꿀때 나누는 기준을 정하기가 어렵다. 따라서 distance measure를 정의하여 통계적으로 음소들을 clustering하게 되면 non-keyword 음성 부분을 보다 효과적으로 모델링할 수 있다. 표 3에는 조음방법에 의해서 non-keyword를 5개의 cluster로 나누었을 때의 keyword 인식률이 나타나있다. 이와 비교하여 표 2에서 cluster 수가 5개인 경우의 keyword 인식률을 보면 두 방법의 성능이 비슷했으며, 이는 조음방법에 따른 분류가 스펙트럼 상에서의 음성의 특징을 포함하고 있다는 것을 보여 준다. 앞으로 clustering된 음소 모델들 사이의 transition에 적절한 penalty를 부가시키면 보다 향상된 인식 성능을 얻을 수 있을 것으로 보인다.

표 1. Keyword spotting baseline 시스템의 keyword 인식률

Keyword 모델	고립단어	문장형태	전 체
단어단위	97.8%	92.2%	95.0%
음소단위	98.9%	86.7%	92.8%

표 2. 통계적 방법에 의한 Non-keyword 모델의 cluster 갯수에 따른 keyword 인식률

Cluster 수	고립단어	문장형태	전 체
1	98.9%	86.7%	92.8%
4	97.8%	90%	93.9%
5	97.8%	88.9%	93.3%
6	100%	93.3%	96.7%
7	100%	93.3%	96.7%
8	98.9%	93.3%	96.1%
9	96.7%	91.1%	93.9%
46	92.2%	83.3%	87.6%

표 3. 여러 가지 non-keyword 모델에 따른 keyword 인식률

Non-keyword 모델	고립단어	문장형태	전 체
10 state를 가지는 단일 모델(Baseline 시스템)	98.9%	86.7%	92.8%
조음방법에 따른 monophone clustering 방법	95.6%	90.0%	92.8%
통계적 방법에 의한 monophone clustering 방법	100%	93.3%	96.7%

V. 결 론

Keyword spotting 기술은 사용자가 자연스러운 연속음성으로 말하더라도 이로부터 미리 주어진 핵심주제어(keyword)들을 검출해 냄으로써 실제적으로 많은 응용분야에 효과적으로 사용될 수 있다. 본 논문에서는 keyword의 변경 및 추가가 용이한 음소모델에 의한 keyword spotting 방식을 구현하고 monophone을 clustering하여 non-keyword 모델로 사용함으로써 그 성능을 향상시키는 연구를 수행하였다. 6개의 keyword를 대상으로 한 화자독립 keyword spotting 실험 결과, 음소모델에 대한 baseline 시스템의 인식률은 고립단어 및 문장형태의 음성데이타에 대해 각각 98.9% 및 86.7%로 나타났다. 그 다음으로 non-keyword 모델 구성 방법으로 단어형태의 단일모델을 사용하는 방법 이외에 조음방법에 의해 monophone을 clustering 하는 방법, 그리고 통계적 방법으로 monophone 을 clustering하는 방법이 검토되었으며, cluster 갯수가 6 또는 7개인 경우의 통계적 clustering 방법의 인식률이 100% 및 93.3%로 가장 우수한 성능을 나타내었다.

현재 keyword 수를 보다 확장시키는 실험과 더불어 후처리 과정을 추가시킴으로써 잘못 인식된 keyword들을 제거시키는 방법에 대한 연구가 진행되고 있으며, 일부 triphone 모델의 경우 훈련용 데이타의 부족으로 적절한 모델링이 이루어지지 못하는 문제를 해결하기 위해 state-tying 방법[9]의 적용이 검토되고 있다

참 고 문 헌

[1] 김 형순, "Keyword Spotting 기술", 한국통신학회지, 제 11권 9호,pp.57-65, 1994.

[2] 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, pp 1870-1878, Nov. 1990.

[3] H. Bourland, B. D'hoore, and J.-M. Boite "Optimizing Recognition and Rejection Performance in Wordspotting Systems," in Proc. IEEE ICASSP, 1994, pp.I-373-376.

[4] R. C. Rose and D. B. Paul, "A Hidden Markov Model Based Keyword Recognition System," in Proc. IEEE ICASSP, 1990, pp.129-132.

[5] M. Weintraub, "Keyword-spotting using SRI's DECIPHER Large Vocabulary Speech Recognition System," in Proc. IEEE ICASSP, 1993, pp.II-463-466.

[6] J. G. Wilpon and L. R. Rabiner, "A Modified K-Means Clustering Algorithm for Use in Isolated Work Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.33, no.3, pp.587-594, June, 1985.

[7] C. H. Lee and L. R. Rabiner, "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," IEEE Trans. Acoust., Speech, Signal Processing, vol.37, no.11, pp.1649-1658, Nov. 1989.

[8] 송 화전, 김 재호, 손 경식, 김 형순, "Keyword Spotting 에서의 후처리과정에 관한 연구", 제 11회 음성통신 및 신호처리 워크샵 논문집, 1994, pp.249-252.

[9] S. J. Young and P. C. Woodland, "State Clustering in Hidden Markov Model-Based Continuous Speech Recognition," Computer Speech & Language, Vol.8, no.4, pp.369-383, Oct. 1994.

본 연구는 한국전자통신연구소 음성언어연구실의 위탁 연구과제 결과의 일부임.