《主 題》

Keyword Spotting 기술

김 형 순
(부산대학교 전자공학과)

■차 례■

Ⅰ. 서 론

음성인식의 궁극적인 목표가 잡음이 있는 실제적인 환경에서 불특정화자가 주제에 구애받지 않고 자연스럽게 발음한 대용량 어휘의 대화체 연속음성을 실시간에 인식 및 이해하는 것이라고 할 때, 선진각국의 오랜 연구노력에도 불구하고 아직까지 이 목표는 달성되지 못하고 있으며 앞으로도 가까운 장래에 실현되지는 않을 전망이다. 따라서, 음성인식에 관한 대부분의 연구들이 이러한 목표 중의 많은 부분에 제약조건을 둔 상태에서 진행되고 있으며, 이에 따라 제한된 응용분야에서는 음성인식이 성공적으로 적용되는 사례들이 속속 등장하고 있다[1]. 지금까지 개발되어 온 음성인식기술을 분류하는 데에는 화자독립성, 인식어휘규모, 문법구조 및 주제 등을 비롯하여 여러가지 기준이 있을 수 있겠지만, 입력음성의 형태에 따라 고립단어인식(isolated word recognition)과 연속음성인식(continuous speech recognition)의 두 가지 부류로 크게 나눌 수 있다. 고립단어인식은 미리 정해진 어휘에 속한 단어들을 한 단어씩 또박또박 띄어 발음한 것을 인식하는 것으로서, 단어 경계가 분명하여 단어연결에 따른 모호성이 없으므로 구현이 용이하고 인식 성능도 우수하지만 사용자가 발음상의 부자연스러움을 감수해야 한다. 이에 반하여 연속음성인식은 문장형태로 자연스럽게 발음한 음성을 인식하는 것으로서, 사용자의 입장에서는 바람직하지만 아직까지 기술수준의 한계로 인하여 매우 제한된 어휘와 문법구조를 갖는 경우를 제외하고는 인식성능이 크게 뒤떨어지는 형편이다.

Keyword spotting(또는 word spotting)은 이들 두 가지 음성인식 방법들 이외의 제 3의 접근 방식으로서, 어휘에 제한없이 자연스럽게 발음한 연속음성으로부터 미리 정해진 특정단어(keyword)들을 검출해 내는 것을 의미한다. 많은 경우 이러한 keyword들은 강세가 주어진 상태에서 발음되거나 충분히 명료하게 발음되며, 연속음성으로 발음된 문장전체를 인식하는 것에 비해서 훨씬 용이한 작업으로 볼 수 있다. 따라서 keyword spotting은 고립단어인식이 지니는 사용자의 불편함과 연속음성인식이 지니는 성능저조의 문제점을 모두 해결할 수 있는 방식으로서, 제한없이 발음된 문장 내에서 핵심어들만 검출해 내면 의미가 통할 수 있는 많은 응용분야에 매우 효과적으로 적용될 수 있다. 이에 따라, 여러 선진국들에서는 이에 대한 연구들이 활발히 진행되고 있으며, 미국 AT&T사 등에서는 keyword spotting 기술을 이용한 전화교환업무의 자동화 서비스를 1992년부터 실용화하여 운영하고 있다.

본 고에서는 이러한 keyword spotting 기술의 전반적인 내용을 개괄적으로 살펴 보고, keyword spotting 기술의 응용 분야를 실제 개발 사례들을 중심으로 기

술하고자 한다.

Ⅱ. Keyword Spotting 방식

2.1 Keyword Spotting 시스템의 기본구성

초기의 keyword spotting 방식들은 dynamic time warping(DTW)이라 불리우는 template matching 방법을 기본으로 하고 있었으나, 최근에 들어서 다른 모든 음성인식 분야와 마찬가지로 음성 신호를 통계적인 모델로서 표현하는 hidden Markov model(HMM) 방법이 주종을 이루고 있다[2][3]. 또한 초기의 keyword spotting 방식들은 keyword에 해당하는 대표적인 패턴들만을 미리 저장해 놓은 다음, 입력음성으로부터 저장된 keyword 패턴들과 잘 부합되는 음성구간을 탐색하여 이 구간에서의 음성신호와 해당 keyword와의 유사도가 미리 정해진 임계치보다 높을 경우 keyword가 검출된 것으로 결정하는 방법을 토대로 하고 있었다. 그러나, filler template(또는 filler 모델)의 개념이 도입된 이후에는 대부분의 keyword spotting 방식이 keyword와 filler에 의한 연결단어인식(connected word recognition)의 구조를 기본으로 하게 되었다[4]. 여기서 filler template(또는 filler 모델)이란 keyword에 해당되지 않는 음성구간(이를 non-keyword라 부름) 및 silence 구간을 대표하는 패턴(또는 모델)으로서 garbage template(또는 garbage 모델)라고도 부른다. 따라서, 본 고에서도 가장 일반적인 keyword spotting 방식으로서 keyword 및 filler에 대한 HMM을 이용한 keyword spotting 방식에 관해 논의하고자 한다.

이러한 keyword spotting 시스템의 일반적 구성도가 그림 1에 나타나 있다. 그림에서 보는 바와 같이 keyword spotting 시스템은 기본적으로 keyword들과 non-keyword 그리고 silence를 각각 HMM으로 모델링하는 것을 구간으로 하고 있다. 이와 같이 non-keyword 및 silence의 모델, 즉, filler 모델을 이용한 keyword spotting 과정은 입력 음성을 일련의 keyword 모델들과 filler 모델들로 표현하여 이들 모델들을 인식 대상 어휘로 하는 연결단어 인식 과정으로 볼 수 있다. 이러한 방식은 keyword 모델만을 사용하는 초기의 keyword spotting 방식의 문제점, 즉, 경우에 따라서는 keyword들끼리 중첩(overlap)되어 인식되는 문제점을 자동적으로 방지하는 효과를 가진다. 다만, 물론 이 방식에서는 filler 모델이 non keyword 음성구간과 silence 구간을 얼마만큼 잘 표현해 주는가에 따라 keyword spotting의 성능이 크게 좌우되게 된다. Filler template들을 HMM으로 모델링할 경우 다수의 화자와 다양한 분맥상에서의 음성들을 통계적인 접근방법에 의해 보다 효과적으로 표현할 수 있으므로 template matching 형태인 DTW 방식에 비해 우수한 성능을 얻을 수 있게 된다.

그림 1에 나타낸 keyword spotting 시스템의 동작을 개략적으로 설명하면 다음과 같다. 그림에서 보는 바와 같이 입력음성이 들어 오면 먼저 음성신호의 전처리 과정인 음성특징추출 과정을 통해 음성신호의 음성학적인 특징을 잘 표현해 주는 특징 파라메터들을 추출한다. Keyword spotting을 비롯하여 음성인식에 널리 사용되는 대표적인 음성 특징 파라메터들로는 LPC(Linear Predictive Coding) cepstrum 계수 및 Mel-frequency cepstrum 계수 등을 들 수 있다. 음성특징 파라메터들이 추출되면 그 다음 단계인 keyword 검출과정에서 미리 구성된 keyword, non-keyword 그리고 silence들에 대한 통계적 모델들을 이용하여 일종의 패턴 매칭 방법에 의해 keyword를 찾아내게 된다. 마지막으로 후처리 과정은 keyword spotting 시스템에서의 오류를 감소시켜 그 성능을 보다 향상시키기 위한

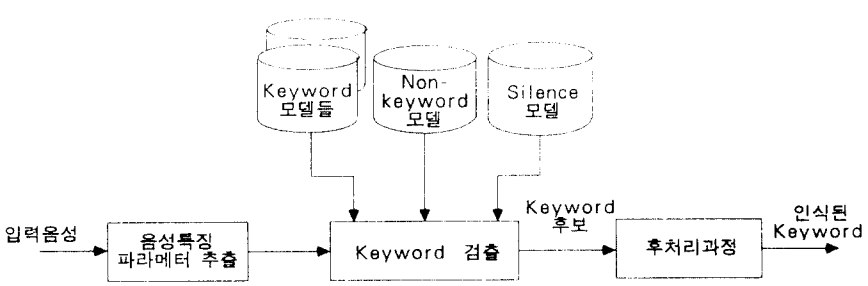

그림 1. Keyword Spotting 시스템의 기본구성도

것으로서, 이미 구해진 keyword 후보들의 신뢰도를 조사하여 keyword로 판단하기 곤란한 것들을 제외시킴으로써 오인식으로 인한 피해를 최소화시킨다.

이상에서 설명한 Keyword spotting 방식에서의 전체적인 HMM network의 구조가 그림 2에 나타나 있다. 그림에서 보는 바와 같이 이 구조는 입력 음성에 어떠한 갯수의 keyword도 포함될 수 있는 형태로 되어 있다. 만약에 하나의 문장에 최대 몇 개의 keyword가 포함될 수 있다는 등의 제약조건이 주어질 수 있다면, 그림 2의 구조를 변형시킴으로써 불필요한 인식 오류를 방지할 수 있다. 예를 들어 그림 3은 하나의 문장마다 하나씩의 keyword가 존재한다는 조건이 주어질 때의 HMM network의 구조를 나타내고 있다.

그림 4는 실제로 keyword spotting이 수행되는 예를 나타내고 있다. "음, 저, 총무과가 몇 번이죠?"라고 입력된 문장 중에서 "총무과"라는 단어가 인식대상 keyword이다. 그림 하단부의 KW, NK 및 S 기호는 각각 keyword, non-keyword 및 silence 모델을 의미하며, 이와 같이 keyword spotting 방식은 입력음성을 가장 잘 표현할 수 있는 keyword들과 non-keyword, 그리고 silence 모델들의 순서열을 찾는 과정을 통해 keyword를 검출하게 된다.

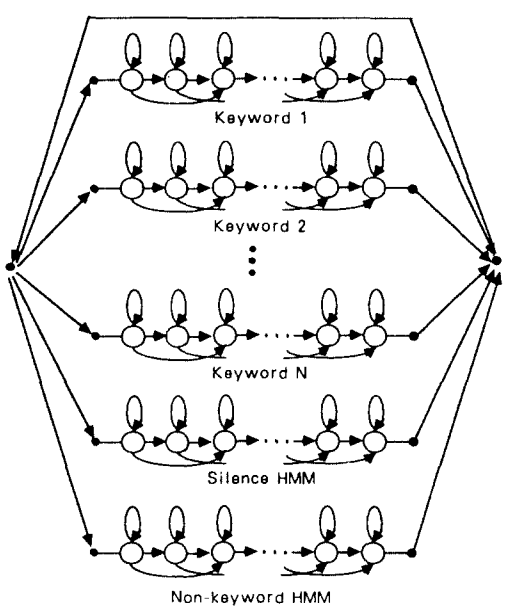

그림 2. Keyword Spotting을 위한 전체 HMM 구조

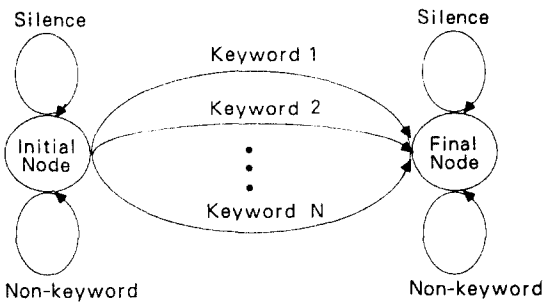

그림 3. 문장당 한개의 keyword가 있을 경우의 keyword spotting 모델의 예

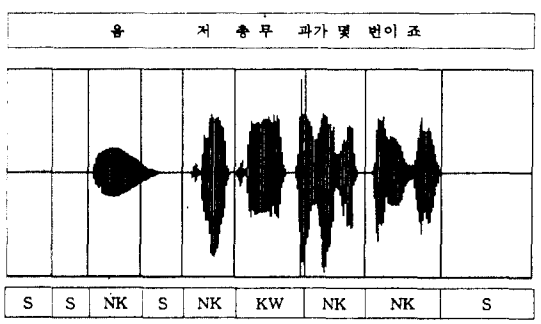

그림 4. Keyword Spotting 과정을 통해 Keyword가 검출되는 예

2.2 Keyword 모델

HMM을 이용한 keyword spotting 방식에서의 keyword 모델들을 구성하는 방법은 크게 단어를 기본단위로 하는 HMM과 음소를 기본으로 하는 HMM의 두 부류로 대별시킬 수 있다. 단어전체를 기본단위로 하는 HMM은 훈련용 음성 데이타에 포함된 keyword들로부터 각각의 keyword를 모델링하는 것으로서, keyword들의 발음사전을 별도로 준비할 필요가 없으며, 인식 대상 keyword들의 갯수가 많지 않고 HMM 모델을 구성하기에 충분한 양의 훈련용 음성 데이타가 확보될 경우 우수한 인식성능을 나타낸다. 이는 단어 전체를 모델링 하기때문에 단어 내에서의 음운변화 또는 상호조음(coarticulation)현상을 모델자체가 포함하고 있기 때문이다. 다만, 이 방식은 기존에 구축된 keyword spotting 시스템에서 keyword의 변경 및 추가가 용이하지 않다는 문제점이 있다. 즉, 이미 구축된 불특정화자 keyword spotting 시스템에서 새로운 keyword를 추가 또는 대치하려면, 그때마다 많은 사람의 음성으로부터 해당 keyword가 포함된 단어 및 문장형

태의 음성 데이타를 수집하여 keyword 모델을 새로 구성해야 하는데, 이를 위해서는 많은 시간과 비용이 소요되게 된다. 따라서, keyword 갯수가 그리 많지 않고 keyword의 변경 및 추가가 필요없는 응용분야에는 단어를 기본단위로 하는 HMM을 이용한 keyword spotting 방식이 구현 및 인식성능 면에서 최적이지만, 그렇지 못한 응용분야에는 바람직하지 못하다.

이에 반하여, 음소 또는 문맥종속형 음소(context dependent phone 또는 triphone)를 기본단위로 하는 HMM 방식에서는 전체 단어들이 미리 정해진 음소들의 연결로 표현된다. 이들 음소들은 여러 단어들에 걸쳐 공통적으로 나타나기 때문에 훈련용 음성 데이타가 이러한 음소들을 모델링할 정도만 준비될 수 있다면 비록 각각의 keyword에 해당하는 음성 데이타는 별로 없더라도 이들 단어들의 발음표기에 따라 음소 모델을 연결하여 단어 모델을 구성할 수 있다. 즉 훈련된 음소 HMM이 구축된다면 새로 keyword를 추가하거나 대치시키는 경우에 이들 음소 HMM으로 모델 구성이 가능하며, 이를 위해 해당 keyword를 포함한 방대한 음성 데이타베이스를 구축해야 할 필요가 없게 된다. 따라서, 음소를 기본단위로 하는 HMM 방법은 소위 text-independent keyword spotting을 가능케 하며, keyword 어휘 변경이 빈번한 경우를 포함한 많은 실용적인 응용분야에 효과적으로 사용될 수 있다. 다만 음소 모델들을 서로 연결하는 과정에서 그 단어 내부의 음운현상들을 잘 모델링하지 못하게 되면 인식성능이 떨어지므로, 문맥종속형 음소모델을 사용하거나 일부의 단어형태의 keyword 음성 데이타로부터 모델을 보완하는 방법 등이 고려될 필요가 있다.

2.3 Non-keyword 모델

이미 언급한 바와 같이 대부분의 keyword spotting 방식들은 keyword 모델과 filler 모델을 사용하는 연결 단어인식 알고리즘을 기반으로 하고 있다. 여기서, filler 모델들은 keyword에 해당되지 않는 음성(non-keyword)구간들과 비음성(silence 또는 배경잡음)구간들을 모델링하는 데에 사용된다. Keyword spotting 방식이 입력된 음성을 keyword 및 filler 모델들의 시간순서열로 매칭시키는 과정을 통해 keyword를 검출하고, 이 때, keyword 모델로는 단어전체 또는 음소를 기본단위로 하는 HMM이 사용되며, keyword spotting 과정에서의 오인식을 감소시키기 위한 한 방편으로서 그림 3에서와 같은 간단한 문법규칙이 적용되기도 한다는 등의 기본 구조가 정해진 다음에는, filler 모델을 어떻게 정의하고 구현할 것인가 하는 점이 keyword spotting 알고리즘들 사이의 가장 큰 차이점으로 부각된다. 그리고, 실제로 filler 모델이 keyword 음성부분을 잡식하지 않으면서 non-keyword 음성부분 및 배경잡음 부분을 얼마만큼 효과적으로 표현해 줄 수 있는가에 따라 keyword spotting 시스템의 성능이 크게 좌우된다. Filler 모델 중에서 silence 또는 배경잡음의 모델은 음성구간이 아닌 부분들을 하나의 모델로 표현해 줄 수 있으므로, 실제 keyword spotting 알고리즘의 관건은 non-keyword 모델의 구현 방법에 주어지게 된다.

Non keyword 모델을 구성하는 방법으로는 빈도가 높은 non-keyword들 각각을 독립된 모델로 구성하는 방법, keyword가 아닌 음성부분을 한 개 또는 여러 개의 대표 모델로 표현하는 방법, 그리고 음소를 비롯한 subword unit들로 non-keyword들을 모델링하는 방법 등이 사용되고 있다. Non-keyword 모델과 관련한 한 연구결과에 따르면, non-keyword 어휘 각각을 HMM으로 모델링하는 방법, 인접음소들 사이의 상호 영향을 고려한 문맥종속형 음소(triphone) 모델을 사용하는 방법, 인접음소 정보를 무시한 음소(monophone) 모델을 사용하는 방법, 그리고 음소 정보마저 무시하고 임의적인 집단화 방법에 의해 다수의 모델을 구성하는 방법에 대해 20개의 keyword를 대상으로 실험한 결과, 이들 4가지 방법들 중에서 처음 3가지 방법의 성능은 거의 비슷했으며, 마지막 방법이 가장 저조한 성능을 보인 것으로 나타났다. 이들 중 첫번째 방법은 방대한 훈련 데이타가 필요할 뿐만 아니라 인식과정에서의 계산량도 매우 많아지는 문제점을 가지며, 두번째 및 세번째 방법에서의 triphone 모델과 monophone 모델의 갯수가 각각 268개와 45개임을 고려할 때, monophone 모델이 가장 효과적인 방법인 것으로 보고되었다[3]. 그러나, 또다른 연구결과에 따르면 non-keyword 음성부분 전체를 하나의 HMM으로 모델링하는 것이 다른 모든 경우에 비해 성능 면에서 뒤떨어지지 않는다는 결과를 얻었다[2]. 실제로 이들 연구결과가 서로 다른 응용분야 및 음성 데이타베이스를 대상으로 하고 있기 때문에 이들로부터 non-keyword 모델에 대한 일반적인 결론을 도출할 수 없으며, 앞으로 보다 효과적인 non-keyword 모델링 방법에 대해 많은 연구가 진행되어야 할 것으로 보인다.

2.4 모델 훈련과정

Keyword spotting을 위해서는 먼저 keyword HMM,

non-keyword HMM 및 silence HMM을 모델링 해야 한다. 이를 위해서는 훈련용 음성 데이타베이스를 keyword와 non-keyword, 그리고 silence 구간으로 분할시켜야 되는데, 여기에는 수동분할(manual segmentation)과 자동분할(automatic segmentation)의 두 가지 방법이 사용될 수 있다. 음성학 전문가에 의한 수동분할은 분할성능 면에서는 우수하나, 불특정화자에 의한 keyword spotting의 성능향상을 위해서는 방대한 음성 데이타베이스를 다루어야 한다는 점을 고려할 때, 많은 시간과 노력을 필요로 하는 수동분할 방법은 바람직하지 못하다.

자동분할 방법은 분할과 훈련이 반복되는 소위 bootstrapping 과정을 통해 구현될 수 있으며, 설명의 편의상 단어를 기본단위로 하는 HMM을 사용하는 경우를 예로 들어 이러한 모델 훈련과정을 설명하기로 한다[5]. 먼저 별도로 준비된 고립단어 형태의 keyword 음성 데이타베이스로부터 1차적인 keyword HMM을 훈련시킨다. 또한 이 데이타베이스에서 자동 음성/비음성 검출과정을 통해 구해진 비음성구간으로부터 1차적인 silence HMM을 훈련시킨다. Non-keyword HMM은 일차적으로 random 데이타로부터 구한다. 이와 같이 일차적인 keyword, non-keyword 및 silence 모델이 구성되면, 이들을 이용해서 문장형태의 keyword spotting 훈련용 음성 데이타베이스로부터 Viterbi decoding에 의해 그림 4에서 보는 바와 같이 자동분할을 행한다. 분할된 데이타들은 이미 keyword, non-keyword 및 silence 구간으로 분류된 상태이므로 이들로부터 keyword, non-keyword 및 silence HMM들의 재훈련과정을 수행하며, 이 과정은 모델들이 수렴상태에 이를 때까지 반복적으로 수행될 수 있다. 이 때, 문장형태의 데이타베이스에 대해서는 훈련중인 문장에 포함된 keyword가 어떤 것이며, 여러 개일 경우 그 순서가 어떻게 되는지에 대한 정보만 제공되면 된다. 이러한 bootstrapping 과정에 의해 음성 데이타베이스의 자동분할과 각각의 HMM의 모델링이 동시에 수행될 수 있으며, 이 과정이 그림 5에 나타나 있다.

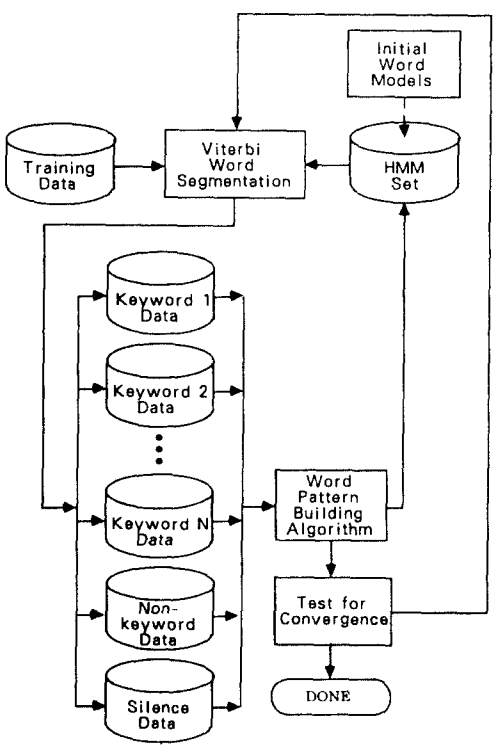

그림 5. Bootstrapping에 의한 모델 훈련과정

2.5 후처리 과정

후처리 과정은 keyword spotting 시스템에서의 오류를 감소시켜 그 성능을 보다 향상시키기 위한 것으로서, keyword spotting 시스템이 검출해 내지 못한 keyword를 후처리 과정에서 찾아낸다는 것은 현실적으로 많은 어려움이 따르기 때문에 여기에서는 이미 구해진 keyword 후보들로부터 잘못 검출된 후보(false alarm)들을 효율적으로 제거하는데에 주안점을 두고 있다. 실제로 후처리 과정은 keyword 후보들의 신뢰도를 적절한 방법으로 조사하여, 신뢰도가 떨어진다고 판단될 경우 keyword 검출결정을 포기하는 방식으로 구현되는 것이 대부분이다. 이러한 과정은 keyword spotting 시스템의 성능이 완벽할 수는 없다는 전제하에서 실용적인 대안을 제시하는 것으로서, 잘못 인식된 keyword에 의해 엉뚱한 응답을 하는 것보다는 다시 발음해줄 것을 요구하거나 차라리 사람이 처리하도록 하는 방법을 취하는 것이 피해를 최소화 시킬 수 있다는 데에 근거를 두고 있다.

이러한 후처리 방식에도 여러 가지 방법이 사용될 수 있으며, 그 중에서 신경회로망을 이용하는 방법, 음성 segment 모델을 이용하는 방법, 변별적 훈련과정을 사용하는 방법, 그리고 keyword 및 filler 모델의 likelihood 비를 이용하는 방법 등이 좋은 성과를 거두고 있다. 여기에서는 그 중의 한 방법으로서 keyword 및 filler 모델의 likelihood 비를 이용하는 방법에 대해 기술하도록 한다[3]. 이 방법은 그림 6에 나타낸 바와

같이 두 가지의 HMM network를 병렬로 사용한다. 그 중 첫번째는 지금까지 설명한 keyword 및 filler network로서 그림 2에 도시된 것과 같으며, 두번째는 keyword 모델없이 filler 모델(즉, non-keyword 모델 및 silence 모델)만으로 구성된 network이다. 따라서, keyword 및 filler network이 입력 문장으로부터 keyword 후보를 검출해 내고 동시에 이 keyword에 해당하는 음성구간에 대한 정보를 filler 모델만으로 구성된 network에 넘겨주면, filler 모델만으로 구성된 network은 이 구간에서의 filler 모델의 likelihood 값을 계산한다. 그 다음 앞서 keyword 및 filler network에서 keyword 검출시에 계산된 likelihood 값과 filler network에서의 filler likelihood 값을 비교하여 그 비가 특정 임계치가 넘지 못하면 그 음성구간에서의 keyword 검출의 신뢰도가 높지 못한 것으로 판단하여 keyword가 검출되지 않은 것으로 간주한다.

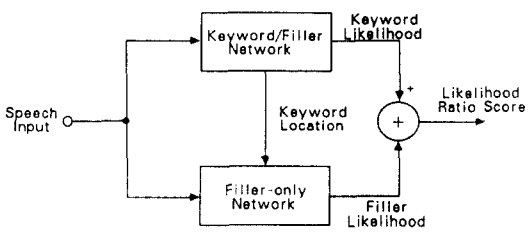

그림 6. Likelihood Ratio Scoring에 의한 후처리 과정

2.6 Keyword spotting 방식의 성능평가기준

Keyword spotting 시스템에서는 3가지 종류의 오류가 발생할 수 있다. 그중 첫번째는 하나의 keyword를 다른 keyword로 오인식하는 경우이고, 두번째는 keyword가 있는데도 이를 검출하지 못하는 경우(miss)이며, 마지막 세번째는 keyword가 아닌 부분을 keyword로 잘못 검출하는 경우(false alarm)이다. 따라서 keyword spotting의 성능은 고립단어인식의 경우처럼 단순히 keyword들에 대한 인식률 또는 오인식률만으로 표현하기는 곤란하다. Keyword spotting 기술은 원래 음성통신 내용의 감시라는 특수한 응용분야를 대상으로 하여 개발되어 왔으며, 이러한 관점에서 keyword spotting 방식의 성능은 일반적으로 단위시간에 keyword 당 몇 개의 false alarm이 발생한 상황(이 값을 false alarm rate라고 부르며 단위는 fa/kw/hr이다)에서 keyword의 검출율이 얼마나 되는가로 표현한다. 여기서 keyword 검출율이란 인식대상 음성에 포함된 전체 keyword의 갯수에 대해 정확하게 검출된 keyword의 수의 비율을 의미한다. 그리고, keyword 검출율을 false alarm rate의 함수로 표시하여 그래프로 그린 것을 receiver operating cure(ROC)라고 부르며, ROC에서 false alarm rate가 0 fa/kw/hr에서 10 fa/kw/hr로 변하는 과정에서의 평균 keyword 검출율을 figure of merit(FOM)이라고 부르는데 이는 ROC의 전체적인 흐름을 단일 수치로 표현한 것이다.

Keyword spotting 기술은 원래 대화내용 감시라는 특수한 응용분야를 바탕으로 연구되어 온 것으로서, 대화내용 감시용 keyword spotting은 고립단어인식 대신 연속음성으로부터 핵심어를 인식하고자 하는, 즉 음성인식과정의 한 수단으로서의 keyword spotting과는 응용분야가 다름으로 인한 약간의 차이점을 지닌다[6]. 물론 이들 두 가지 응용분야는 그 기본적인 목적상 불특정화자에 의한 연속음성으로부터의 단어인식을 전제로 한다는 점에서 동일한 내용을 다루고 있지만, 감시용 keyword spotting은 음성인식이 진행되는 사실을 알지 못하는 비협조적인 화자를 대상으로 하는 반면에, 음성인식의 수단으로서의 keyword spotting에서는 많은 경우 인식이 되기를 바라는 협조적인 화자를 대상으로 한다. 감시용 keyword spotting과 음성인식의 수단으로서의 keyword spotting의 보다 중요한 차이점은 keyword spotting과정에서의 오류가 응용분야에 미치는 영향이 다르다는 점이다. 감시용 keyword spotting의 경우, 문장에 포함된 keyword를 검출하지 못하는 것은 치명적이지만, 어떤 keyword를 다른 keyword로 오인식하는 것은 결국 operator가 이를 청취하는 과정에서 처리할 수 있으며, keyword가 아닌 부분을 keyword로 잘못 검출하더라도 operator의 노력낭비를 가져오는 정도의 피해가 전부이다. 그러나, keyword spotting을 음성인식에 적용할 경우에는 keyword spotting을 검출못하게 되더라도 이에 대한 반응을 할 수 없는 정도로 끝나지만, keyword를 오인식하거나 keyword가 아닌 부분을 keyword로 잘못 검출하게 되면 예기치 못한 결과를 초래할 수도 있다. 다시 말해서, 응용분야가 어떤 것인가에 따라서 keyword spotting의 각종 오류가 미치는 영향이 달라질 수 있으며, keyword spotting의 성능도 실제 용도에 따라 각각의 오류에 대한 가중치를 조절하여 표현할 필요가 있다. 극단적인 예로서 입력된 문장에 단 하나의 keyword가 포함되어 있다는 제한조건이 주어질 수 있는 응용분야의 경우, 그림 3에 나타낸 바와 같이 반드

시 하나의 keyword가 검출될 수 있도록 문법적 제약을 들 수 있으며 이때의 keyword spotting의 성능은 단순히 keyword 인식률만으로 표현될 수 있다.

Ⅲ. Keyword Spotting 기술의 응용

3.1 개 요

Keyword spotting 기술은 원래 음성통신에 대한 감시용의 목적으로 개발되어 온 것으로 음성인식 분야의 연구 초기부터 지속적으로 연구가 수행되어 왔으며, 1977년에 발표된 Dialog System 사의 keyword spotting 시스템은 음성 메세지로부터 'Kissinger'라는 하나의 단어를 검출해내는 실험을 수행한 것으로 보고되고 있다[7]. 그러나, 지금까지 개발되어온 음성인식기술을 실제 응용분야에 적용하려는 시도들이 본격적으로 추진됨과 더불어 음성인식의 한 방편으로서 keyword spotting 기술의 유용성이 부각됨에 따라 이에 대한 활발한 연구가 진행되고 있으며, 그 결과로 다양한 응용분야에 keyword spotting 기술이 사용되기 시작하고 있다.

사용자가 자연스럽게 발음한 연속음성으로부터 핵심주제어를 추출해낼 수 있는 keyword spotting 기술의 응용분야는 매우 넓다. 즉, 음성 명령어에 의한 컴퓨터 및 각종 기기의 작동을 비롯하여 기존에 고립단어 인식기술이 적용되어 온 모든 응용분야에 무제한적인 연속음성의 입력이 가능해지게 된다. 특히, 가장 널리 보급된 정보단말기인 전화를 이용하여 컴퓨터가 알아들도록 말하는 데에 익숙하지 않은 일반인들이 자연스러운 음성입력에 의해 여러가지 정보 서비스를 제공받을 수 있게 됨으로써, 정보화의 촉진에 크게 기여할 수 있다. 이러한 서비스의 예로는 전화교환 서비스, 전화번호 안내 서비스, 증권정보를 비롯한 각종 생활정보 서비스 등을 들 수 있으며, 선진국들의 여러 전화사업자들이 현재 이러한 서비스를 이미 시행하고 있다[1][8].

또한 keyword spotting 기술은 일반적인 연속음성인식의 전처리기 또는 보조처리기로서 효과적으로 활용될 수 있다. 주제에 상관없이 자연스러운 대화체로 발음한 연속음성의 인식은 현재의 기술수준으로 달성되기 어렵다는 판단에 따라, 대부분의 연속음성인식 기술은 항공여행이나 호텔예약 등의 특정 주제 하에서 다루어지고 있으며 문장구조에도 제한을 두는 경우가 많다. 그러나, 실제적으로 많은 응용분야들이 제한적인 단일 주제로만 국한시키기는 곤란한 실정이며, 여러 주제를 함께 다룰 경우 주제범위가 커지는 데에 따른 탐색(search) 영역의 증가가 인식성능에 큰 영향을 미치게 된다. 따라서, keyword spotting 기술을 통해 입력된 연속음성으로부터 핵심주제어를 자동추출하게 되면 이를 이용하여 주제범위를 세분화하여 불필요한 탐색을 줄일 수 있으며, 문법적으로 모호하여 연속음성인식이 다루지 못하는 대화체 음성에 대해서도 최소한의 대응이 가능해진다.

본 고에서는 이와 같이 다양한 keyword spotting 기술의 응용분야들 중에서 실제 구현 사례들을 중심으로 한 몇 가지 응용예들을 살펴 보기로 한다.

3.2 AT&T VRCP(Voice Recognition Call Processing) 서비스

Keyword spotting 기술이 실제 서비스에 응용된 대표적인 예로 미국 AT&T 사의 VRCP(Voice Recognition Call Processing) 서비스를 들 수 있다[8]. 이 서비스는 0번으로 시작되는 교환원 지원 통화(operator assisted calls)를 자동화시킨 것으로서, 사용자는 음성으로 다섯 가지 서비스 메뉴 중 하나를 선택하도록 구성되어 있다. 처음에는 이 서비스를 위해 불특정화자에 의한 고립단어 인식기술을 사용하려고 했지만, 다수의 전화 사용자를 대상으로 한 현장실험결과에 따르면, 음성안내를 통해 고립단어만을 발음하도록 요구해도 20 퍼센트 가량의 사용자들이 이를 무시하고 문장형태로 발음한다는 사실이 조사되었다. 따라서, 이들에 대해서도 서비스를 지원하기 위해서는 무제한적인 연속음성으로부터 keyword를 찾아내는 keyword spotting 기술이 필수적으로 요청되게 되었다. 이 서비스에서는 각각의 서비스 메뉴를 위해 callect, calling card, person-to-person, third number 및 operator의 다섯 개의 keyword를 선정하였는데, 이들은 각각 수신자 요금부담 통화(callect call), 전화카드 통화(calling card call), 수신자 지정 통화(person-to-person call), 제삼자 요금부담 통화(bill-to-third-party call), 그리고 교환원 직접연결을 의미한다.

이 서비스는 1992년에 개시되었는데 그 해에만도 매일 평균 1160만 호의 서비스 통화를 처리했으며, 이는 1년에 42억 호의 통화를 소화해내는 규모로서 단일 서비스 만으로 연간 3억 달러의 비용절감을 하는 효과에 해당하며 매년 그 규모가 증가되는 추세에 있다[8]. 실제로 이 서비스에 사용된 keyword spotting 시스템은 단어를 기본단위로 하는 HMM을 사용하고 있으며 고립단어 입력의 경우 99.3%, 그리고 문장 당

keyword가 하나씩 포함된 연속음성 입력의 경우 95.1%의 keyword 인식률을 나타내었다.

3.3 전화번호 안내시스템

역시 미국의 전화회사인 NYNEX사에서는 우리나라의 114 서비스에 해당되는 전화번호 안내(directory assistance) 서비스를 자동화하기 위한 사업의 일환으로, 전화번호를 알고자 하는 곳이 속한 도시명을 keyword spotting 기술로 인식하는 서비스를 시험 중에 있다[9]. 이 서비스는 교환원에 의한 전화번호 안내의 전 단계로서 사용자에게 서비스를 원하는 도시명을 컴퓨터 음성으로 물어본 후 사용자가 문장형태로 대답하더라도 이를 인식하여 해당 도시를 담당하는 교환원에게 넘겨주는 방식으로 구성되어 있다.

한국전자통신연구소 자동통역연구실의 위탁연구과제로 부산대학교에서 개발중인 keyword spotting 시스템은 간단한 형태의 부서 전화번호 안내시스템이다. 이 시스템은 사용자가 특정 부서에 대한 전화번호를 문의하면 문장형태의 음성으로부터 부서명을 인식하여 이에 대한 적절한 응답을 하게 된다. 총 22개의 부서명을 인식대상 keyword로 하고 있으며, 6개의 keyword를 대상으로 한 baseline 시스템의 경우 불특정화자의 연속음성 입력에 대해 92.2%의 keyword 인식률을 얻었다.

3.4 Voice GREP 기능

여러 개의 문서 파일들 중에서 어떤 특정 내용을 검색하기 위해서는 찾고자 하는 내용의 keyword에 해당하는 단어가 그 파일들 중의 어느 위치에 나타나는지를 확인 한 다음, 이러한 keyword 부근을 살펴봄으로써 원하는 내용이 어디에 들어 있는지를 쉽게 확인할 수 있다. 실제로 Unix 운영체제에서의 grep 명령어나 Norton Utility의 ts(text search) 명령어 등이 이와 같이 파일들 중에서 특정 문자열을 찾아내는 기능을 수행해 준다. 그러나, 방대한 양의 음성 메세지 파일 중에서 원하는 내용이 들어있는 위치를 찾기 위해서는 일반적으로 음성 메세지 파일 전체를 다 들어보는 수 밖에 없다. 이 때, keyword spotting 기술은 찾고자 하는 내용의 keyword가 들어있는 위치를 검출해 냄으로써 음성 메세지 파일에서의 voice GREP 명령어 역할을 수행해 줄 수 있다. Voice GREP 기능은 음성 사서함 메세지들 중에서 원하는 내용만을 선별적으로 검색하거나, 비디오 테이프에 수록된 강좌의 특정부분을 발췌하거나, 연설 내용을 편집한다는 등의 찾아내다듬지 하는 등의 다양한 응용분야에 적용될 수 있다. 이러한 응용분야들은 경우에 따라 keyword를 빈번하게 교체할 수 있어야 하므로, 음소 등을 기본단위로 하는 text-independent keyword spotting 기술이 필수적으로 요청된다. 미국 Xerox사에서는 특정화자 및 화자적응 방식으로 음성 메세지를 탐색하는 voice GREP 기능을 구현하였는데[10], 연설이나 강좌 등과 같이 특정인의 음성 메세지가 상당히 긴 시간동안 지속되는 응용분야에서는 이와 같이 화자적응 방식을 도입함으로써 일반적인 불특정화자를 대상으로하는 경우보다 인식성능을 높일 수 있다. 실제적인 응용을 위해서는 실시간보다 훨씬 빠른 처리속도를 갖도록 계산량을 감축하기 위한 방안 등이 심도깊게 검토될 필요가 있다. 대부분의 음성 사서함 시스템들이 저장용량을 줄이기 위해 음성신호를 압축하여 저장하기 때문에 압축/재생된 음성에 대한 성능평가들도 시도되고 있다.

3.5 Topic Identification

Topic identification이란 음성 메세지가 미리 정해진 몇 가지 가능한 주제들 중에서 어느 범주에 속하는 것인지를 자동적으로 식별해 내는 작업을 의미하며, 이는 앞서 언급한 음성 사서함 메세지에서 keyword를 찾아내는 단계에서 더 발전한 것이다. Topic identification 과정은 주제식별에 효과적인 keyword 주제를 선정하여 이를 keyword spotting 기술로 검출한 다음 이들 keyword들로부터 주제를 추정하는 작업으로 이루어진다. 이 경우 단지 몇 개의 keyword가 나타났다고 해서 주제를 단정하기는 곤란하므로 일반적으로 인식대상 keyword 규모를 크게 해서 통계적인 접근방법이 사용된다.

미국 BBN사에서는 Switchboard라 불리우는 전화음성 데이타베이스로부터 공기오염, 음악, 범죄, 차량구매, 공공 서비스 등 10개의 주제를 추출하여 주제식별 시험을 하였는데 750개의 keyword를 대상으로 한 keyword spotting 기술을 적용하여 최고 90.9%의 식별율을 얻었다. 이 결과는 문자형태의 정보를 대상으로 하였을 경우의 식별율 98.8%에는 크게 뒤지는 것이지만 음성 메세지의 주제식별 가능성을 보여주고 있다[11].

Ⅳ. 결 론

앞으로의 음성인식기술 개발은 크게 두 가지 방향

으로 나누어 질 수 있을 것이다. 그 중 첫번째는 음성인식의 궁극적인 목표인 대화체 연속음성의 인식을 추구하는 방향으로서 이를 위해서는 대화체 음성언어의 처리에 큰 비중이 두어져야 할 것으로 보인다. 두번째는 현재의 기술수준을 토대로 음성인식기술의 실용화를 추구하는 방향으로서 잡음환경에서의 음성인식 기술, 새로운 화자에 대한 적응기술, 그리고 사용자의 편리함을 고려한 keyword spotting 기술 등이 이러한 범주에 포함될 것이다.

Keyword spotting 기술은 연속음성으로부터 핵심주제어들을 검출해 냄으로써 인간과 컴퓨터 사이의 (비록 최소한의 수준이기는 하지만) 자연스러운 의사소통을 가능케 해 준다는 점에서 사용자의 편리함이 강조되는 추세와 더불어 음성인식의 한 방법으로서 그 역할의 중요성이 점차 증대되고 있다. 이미 살펴 본 바와 같이 keyword spotting 기술은 많은 실용적인 응용분야들에 효과적으로 사용될 수 있으며, 실제로 keyword spotting 기술이 더욱 폭넓은 분야에 사용되기 위해서는 효과적인 non-keyword 모델링 방법 및 후처리 방법을 비롯하여 보다 우수한 성능을 얻기 위한 많은 연구들이 이루어져야 할 것이다. 이에 따라 여러 선진국들에서 이 분야에 대한 활발한 연구가 진행되고 있으며, 비록 늦은 감은 있지만 국내에서도 keyword spotting에 관한 연구가 시작되고 있음은 다행한 일이라 여겨진다. 앞으로 우리 말을 이용하는 응용분야마저도 외국의 기술에 종속되지 않기 위해서는 keyword spotting을 비롯한 음성인식기술의 개발 및 실용화에 더 많은 관심과 노력이 기울여져야 할 것으로 판단된다.

참 고 문 헌

1. D. B. Roe and J. G. Wilpon, "Whither speech recognition : the next 25 years," IEEE Communication Magazine, Vol.31, No.11, pp.54-62, Nov. 1993.
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. R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," in Proc. IEEE ICASSP, 1990, pp.129-132.
4. A. L. Higgins and R. E. Wohlford, "Keyword recognition using template concatenation," in Proc. IEEE ICASSP, pp.1233-1236, 1985.
5. L. R. Rabiner, J. G. Wilpon and B. H. Juang, "A segmental k-means training procedure for connected word recognition based on whole word refrence patterns," AT&T Tech. J., vol.65, no.3, pp.21-31, May 1986.
6. J. R. Wilpon, L. G. Miller, and P. Modi, "Improvments and applications for keyword recognition using hidden Markov modeling techniques," in Proc. IEEE ICASSP, 1991, pp.309-312.
7. W. A. Lea, "Speech recognition : past, present, and future," in Trends in Speech Recognition, W. A. Lea, Ed., Prentice-Hall, Inc., Englewood Cliffs, pp.39-98, 1980.
8. L. R. Rabiner, "Applications of Voice Processing to Telecommunications," Proc. IEEE, vol.82, no.2, pp. 199-228, Feb. 1994.
9. B. Chigier, "Rejection and keyword spotting algorithm for a directory assistance city name recognition application," in Proc. IEEE ICASSP, 1992, pp. II-93-96.
10. L. D. Wilcox and M. A. Bush, "Training and search algorithms for an interactive wordspotting system," in Proc. IEEE ICASSP, 1992, pp.II-97-100.
11. J. McDonough et al., "Approaches to topic identification on the switchboard corpus," in Proc. IEEE ICASSP, 1994, pp.I-385-I-388.

본 고의 내용은 한국전자통신연구소 자동통역연구실의 위탁연구과제 수행내용의 일부를 토대로 작성된 것입니다.

김 형 순

- 1960년 8월 21일생
- 1983년 : 서울대학교 전자공학과 졸업 (공학사)
- 1984년 : 한국과학기술원 전기 및 전자공학과 석사과정 (박사과정 조기진학)
- 1989년 : 한국과학기술원 전기 및 전자공학과 졸업 (공학박사)
- 1987년 ~ 1992년 : 디지콤 정보통신연구소 연구부장
- 1992년 ~ 현재 : 부산대학교 전자공학과 조교수
- 주관심분야 : 음성인식, 음성합성, 디지탈 통신