500+ 단어 단독어 음성 인식 시스템

이 강성
광운대학교 컴퓨터공학과

500+ words Isolated-word Speech Recognition System

Lee, Gang Sung
Computer Engineering Dept., Kwangwoon Univ.
Email : gslee111@daisy.kwangwoon.ac.kr

Abstract

This paper describes an overview of the system designed for 500-word speech recognition. The system is based on the triphone models and uses Dynamic Multisection(DMS) technique for pattern matching. The system is very flexible in the sense of the word-dictionary which is changable spontaneously without any training. The vocabulary selected for the experiments is 561 words of province names, district names of Seoul and Pusan. The experimental results which will be shown here are preliminary because only one speaker was involved in the experiment. But the result is satisfactory when we see the performance is 95.1%. The system is designed on the Windows-95 and works in realtime on the Pentium-133 computer.

1. 서론

소규모 어휘 (수 십 단어)의 음성 인식 시스템은 대부분 단어 단위의 모델을 사용하였다. 그러나 어휘의 수가 중규모(수백)혹은 대규모(수천)로 증가할 때 단어 단위의 음성 인식은 필요한 기억 장치나 계산량으로 볼 때 어렵다. 따라서 많은 시스템들이 음소 모델과 같은 작은 단위를 이용하여 단어나 연속어를 인식 해왔고 좋은 성과를 얻었다[1-2]. 이러한 부 단어 단위의 접근 방법의 장점은 단순히 계산량이나 기억 용량을 줄이는 것 외에도 인식 대상 어휘를 자유스럽게 바꿀 수 있는데 있다. 음소가 기본 모델이므로 텍스트의 사전만 구성하면 학습 없이도 원하는 임의의 단어를 인식할 수 있는 매력은 아주 크다. 기존의 많은 음성 인식 시스템들은 인식 대상 어휘를 학습을 시켜야 하거나 화자 독립인 경우도 지정된 어휘 이외에는 인식 할 수 없었다. 그러나 이 어휘 독립 기능은 그때 그때의 상황에 따라 필요한 어휘를 즉각적으로 학습 없이 사전에 등록할 수 있기 때문에 응용 프로그램 메뉴의 인식, PC통신의 메뉴의 인식, 웹 링크 등등의 가변 어휘를 필요로 하는 응용에 적용할 수 있다.

임의의 어휘를 인식할 수 있는 본 어휘 독립 단어 음성 인식 시스템은 DMS(Dynamic multisection)기법[3]을 기반으로 구성되었으며 음소 단위 모델은 트라이폰(triphone)을 사용하였다. 사용 어휘는 전국의 대도시 명, 도명, 서울시와 부산시의 구명, 동명으로 총 561개 단어이지만 중복 단어를 제외하면 총 542개 단어이다. 여기서 동명을 주요 대상 어휘로 선정한 이유는 특별한 응용 목적이 있어서라기 보다는, 인식 대상어를 쉽게 확장할 수 있다는 것(1000단어 혹은 그 이상)과 동명은 많은 유사한 단어를 포함하고 있다는 것이다. 따라서 인식 시스템을 평가하기에 좋은 어휘 집합으로 보인다. 단, 특정한 음절('동')이 지나치게 많이 발생하는 것이 어휘 구성상에 문제라 하겠다.

검색 시간을 줄이기 위하여 시간 동기 (time synchronous bean search) 동적 프로그래밍을 통한 가지치기와 빔검색(beam search)을 사용하였으며, 이를 위해서 트리 사전(tree lexicon)을 구성하였다[4]. 트리 사전은 단어를 저장한 텍스트 선형 사전(linear lexicon)으로부터 자동 구성되며 적절한 음운 규칙에 의하여 각 음소는 트라이폰으로 변환된다.

그림 1에 인식 시스템의 블록 다이아그램을 보인다. 임의의 아스키 문자로 된 단어 사전이 입력되면 문자-음성 변환 규칙에 따라 문자를 발음 나는 음소의 열로 변환한다. 이 선형 발음 단어 사전은 사전 구조 변환기에서 트라이폰으로 전환되며 트라이폰이 존재하지 않을 경우는 가장 가까운 트라이폰으로 대치된다. 그리고 선형 구조가 효과적인 검색을 할 수 있는 트리 구조로 변환된다. 패턴 비교기에서는 이 트리 사전과 음소사전을 이용하여 단어 검색을 하며 복수개의 후보를 출력한다. 최종적으로 결정 규칙을 통해서 선택된 결과를 낸다.

2. 인식 시스템

본 인식 시스템은 음소 인식을 기반으로 하는 화자 종속, 어휘 독립 단어 인식 시스템이다. 임의의 어휘에 대해서 인식

할 수 있다. 500개 이상의 단어도 인식 가능하나 현재 실험 및 평가된 어휘 수는 500단어 정도이다. 단어 수가 증가한다면 인식율과 인식 시간을 줄이기 위한 알고리즘이 개발되야 할 것이다. 음소 모델은 트라이폰(triphone)이며 실험에서 사용된 542개 단어인 경우, 트라이폰의 수는 1103개이다. 각 트라이폰은 3개의 상태를 갖는 DMS로 모델링 되었다.

그림 1의 각 처리기를 간략하게 설명한다. '문자 음성 변환기'는 문법에 맞춰서 기술된 텍스트 단어를 소리나는 대로 바꾼다. '사전 구조 변환기'는 선형으로 되어있는 사전 구조를 초기의 같은 음소를 공유할 수 있는 트리 구조로 변환한다. '특징 벡터 추출'은 12차의 LPC 멜 캡스트럼 개수를 10kHz, 16비트로 샘플링 된 샘플들로부터 구한다. '패턴 비교기'는 시간 동기적으로 트리 사전에서 분기가능한 음소를 구하고 음소 DB에서 실제 모델을 가져와서 각 경로를 유지하면서 동적 프로그래밍 방법으로 경로의 비용을 계산한다. 최후에 복수개의 후보가 나타날 수 있으며 그 결과들은 결정규칙으로 넘어간다. '결정 규칙'에서는 몇 개의 가장 가능성 있는 후보를 출력한다.

'문자-음성 변환기'와 '사전 구조 변환기'에 대해서는 3절에, 트리 검색에 관해서는 4절에 설명한다.

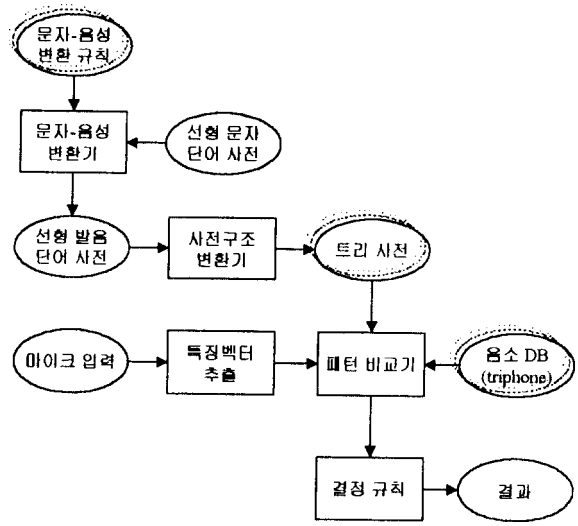

그림 1 인식 시스템 블록 다이아그램

3. 단어 사전

그림 1에서 '선형 문자 단어 사전'은 단순히 단어 텍스트의 모음이다. 일부를 보이면 다음과 같다.

서울 부산 인천 광주

위와 같은 단어는 '문자-음성 변환 규칙'을 이용하여 발음나는 문자열로 변환된다. 예를 들면 '경상북도' ⇨ '경상북또', '개포동' ⇨ '갭포동'등과 같다. '문자-음성 변환 규칙'의 한 예는 다음과 같다.

[__K __T __P] K__ => * KK__

위 규칙은 종성 'ㄱ', 'ㄷ', 'ㅂ'다음에 초성 'ㄱ'이 나올 경우에는 따라나오는 초성 'ㄱ'이 'ㄲ'으로 변한다는 의미이다. 단어 '학교'는 '학꾜'로 변환되는 것과 같다.

각 규칙의 기술은 교육부에서 발간한 '국어어문규정집'을 참고하였다[5]. 모든 단어가 규칙에 따라서 올바르게 변환되는 것을 보장할 수 없으므로 규칙에 어긋나는 단어들은 예외 처리 섹션에 기술되어 우선적으로 변환이 적용된다.

'문자-음성 변환기'를 통과한 사전은 발음 음소로 기술된 단어 사전이다. 선형 구조는 아래와 같다.

강남	ㄱ ㅏ ㅇ ㄴ ㅏ ㅁ
강동	ㄱ ㅏ ㅇ ㄷ ㅗ ㅇ
강북	ㄱ ㅏ ㅇ ㅂ ㅜ ㄱ
강서	ㄱ ㅏ ㅇ ㅅ ㅓ

이러한 구조에서는 두 세 개의 단어가 같은 초기 음소를 공유한다고 해도 반복하여 패턴을 매칭해야 한다. 따라서 효율을 살리기 위하여 사전 구조 변환기는 그림 2와 같이 트리 구조로 변환된다. 루트는 단어의 시작이며 단말 노드는 단어의 끝이다.

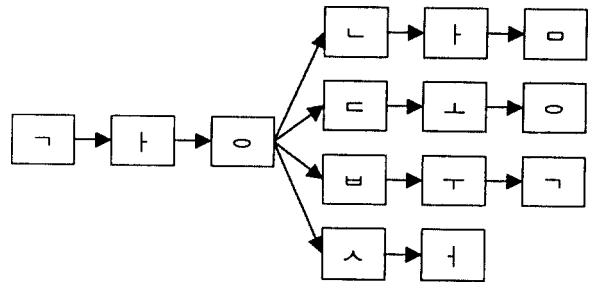

그림 2. 음소 트리 사전 예

그러나 높은 인식율을 내려면 단순한 단음소(monophone) 모델이 아니라 트라이폰 모델을 사용해야 한다. 트라이폰은 일반적으로 더 많은 노드와 아크(arc)를 사용한다. 트라이폰 트리 구조로 변환된 모양을 그림 3에 보인다. 그림 3에서 'A B C'는 좌로 'A', 우로 'C'가 올 경우의 음소 'B'를 의미한다.

트리 구조는 같은 음소로 시작하는 단어에 대해서 같은 음소에 대한 중복 계산을 방지해준다. 선형 구조의 사전에서의 노드의 수가 4288개인데 반하여, 트라이폰 트리에서의 노드 수는 3540개이다. 어휘의 수가 증가하면 할수록 노드수의 차이는 커지게 마련이다. 표1에 각 레벨에서의 노드의 수를 보인다.

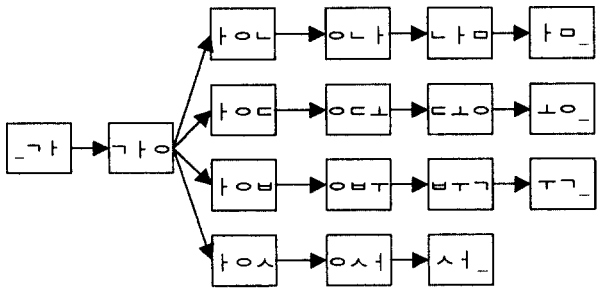

그림 3. 트라이폰 트리 사전 예

표 1. 각 레벨에서의 노드수

레벨	1	2	3	4	5	≥6
선형 사전 노드수	118	322	468	515	529	1586
트리 사전 노드수	542	540	540	540	537	1587

4. 트리 검색

트리 사전의 검색은 시간 동기(time-synchronous) 동적 프로그래밍 기법으로 수행된다. 즉, 한 경로를 따라 계산이 끝난 후에 다른 경로를 계산하는 것이 아니라 현재 시간 t에 대한 입력 프레임 f_t가 주어져 있을 때 모든 가능한 경로에 대해서 동시에 입력 프레임에 대한 거리를 계산하는 것이다. 이렇게 함으로써 상대적으로 가능성이 낮은 경로들에 대해서 가지치기(pruning)를 해 나갈 수 있다. 또한 빔검색을 하여 계산량을 줄인다. 이로써 실시간 인식기 구현이 가능하였다.

5. 실험 및 고찰

트라이폰을 구성하기 위해서 사용한 어휘는 기본적인 단모음, 자음, 전국의 대도시명, 도명, 서울시의 구명, 동명, 부산시의 구명 및 동명으로 정했다. 동명 중에서도 '역삼1동', '역삼2동'등과 같은 것은 '역삼동' 하나로 통합했고 '을지로1가'등은 '을지로'로 통합하였다. 음소 학습을 위해서 총 561개의 단어를 남성 화자 1인이 1회 녹음하였다. 10kHz, 16비트로 녹음된 샘플들은 12차의 LPC 멜켑스트럼 계수로 변환되었다.

사용된 컴퓨터는 인텔 펜티엄-133을 장착한 IBM-PC 계열이고, 운영체제는 윈도우95이다. 음성은 사운드 블래스터 카드로 수집되었다.

트라이폰 모델을 구하기 위한 세그먼트 작업은 초기에 단모음과 몇몇 자음의 기본적인 음소에 대한 구분 정보를 수동으로 입력해 주었고, 나머지는 세그멘탈 k-means 기법을 기반으로 한 자동 세그멘테이션 알고리즘으로 수행하게 하였다. 세그먼트 정보는 이 후에 툴을 이용하여 수동으로 약간 교정되었다. 이렇게 교정하는 이유는 학습 어휘와 인식 대상 단어가 일치한다면 자동 세그멘테이션이라도 무방하지만, 대상 어휘가 다르다면 발생할 수 있는 잘못된 세그먼트 정보로 인한 오류를 조금이라도 줄이기 위해서이다. 총 1103가지의 트라이폰이 추출 되었다.

5.1. 동일한 어휘에 대한 인식 실험

학습에 사용했던 동일한 어휘를 인식 실험에 사용한다. 지역명은 총 561개지만 중복되는 단어 19개를 제외하면 542개이다. 542개 단어를 음소로 분리해서 그 수를 계산하면, 음소의 수가 4303개 나오며, 각 음소당 3개의 상태를 갖는 DMS모델이 구성되었으므로 전체 상태 수는 12909개이다. 542개 단어 구성을 위해 사용된 트라이폰의 개수는 1066개이다. 학습을 위하여 561개 단어를 2회 발성하여 인식 실험한 결과 95.1%가 나왔다.

5.2. 다른 어휘에 대한 인식 실험

음소 모델 작성에 전혀 사용되지 않은 어휘 120개를 선정하였다. 모델에서 지원하지 않는 트라이폰인 경우에는 음소간 거리 표를 작성하여 가장 유사한 트라이폰을 대치하는 단순한 알고리즘을 사용하였다. 실험결과 전체 1129개의 음소 중 120개의 트라이폰 음소가 대치되었다. 이 어휘에 대하여 동일한 화자 1인이 120단어를 1회씩 녹음하여 실험한 결과 89.1%의 인식율을 보였다.

5.3. 인식 시간에 대하여

위에서 수행된 실험에 대하여 인식 시간을 계산하였다. 입력 음성이 길 경우에는 인식 시간이 길게 걸리고 짧을 경우에는 인식 시간이 짧게 걸리는 것을 쉽게 예측할 수 있다. 그러나 이러한 절대적인 인식 시간이 중요하지는 않다. 왜냐하면 사람도 긴 말의 경우에는 그 만큼 이해 하는 시간이 필요하기 때문이다. 따라서 인식 시간을 측정하기 위해서 다음과 같은 Realtime 단위를 사용한다.

$$Rt = \text{인식 시간} / \text{입력 음성의 길이} \qquad (1)$$

이 수식은 입력 음성의 길이와 인식 시간이 같다면 실시간 인식이라 정의한다. 만일 음성을 입력받으면서 동시에 인식 연산을 수행한다면 Rt = 1의 의미는 말이 끝남과 동시에 인식 결과를 볼 수 있다는 뜻이다. 음성을 입력받고 나서 인식하느냐 받는 중에 인식하느냐는 프로그래밍 테크닉으로써 인식 알고리즘 자체의 성능과는 무관하다. 음성 획득과 함께 인식 연산을 병행적(concurrently)으로 혹은 병렬적(parallel)으로 수행한다면 $Rt \geq 1$를 만족한다. $Rt > 0$인 조건하에서 식(1)을 이용하여 앞의 두 실험에서 평가한 결과 어휘가 동일할 경우 평균 Rt = 0.93, 분산=0.52, 어휘가 다를 경우 평균 Rt = 1.01, 분산=0.76임을 보았다. 어휘수가 작음에도 평균 인식 시간이 더 증가한 것으로 보아 어휘가 다를 경우에는 검색 범위가 훨씬 더 넓어짐을 알 수 있다.

6. 결론

본 인식 시스템은 음소 인식을 기반으로 하는 화자 종속, 단어 인식 시스템이다. 트라이폰을 기본 인식 단위로 사용하며, 모델링은 DMS 기법이 사용되었다. 대상 어휘는 학습 없이 임의로 바꿀 수 있으며, 현재 500단어 수준으로 인식기를 평가하였다. 인식율은 학습 어휘와 인식 어휘가 동일한 범위일 때 95.1%를 보였고, 어휘가 전혀 다를 때 89.1%를 보였다.

사용한 트라이폰의 수는 1103개로서 이는 전체 필요한 수에 아주 미치니 못하는 수차이지만 이 트라이폰으로 구성된 어휘가 인식대상이 일 때는 인식율이 좋았다. 그러나 트라이폰이 없는 어휘가 입력되었을 때는 인식율은 상대적으로 많이 저하되었는데 이는 트라이폰 모델을 갖추지 못한데서 기인다. 인식 시간은 인식 시간에 대한 발음 시간의 비로 계산되며 어휘가 동일하고 단어 수가 561일 경우에 Rt=0.93, 어휘가 다르고 단어수가 120일 경우에 Rt=1.01이 나옴으로 실시간 인식임을 보였다.

앞으로는 음소 학습 어휘 집합을 음소 균형을 이룬 것으로 늘리고, 트라이폰을 확대하며, 음소 치환 기법을 개발할 것이다. 또한 어휘 수도 500단어 수준에서 1000단이 수준으로 확대 실험할 예정이다. 이에 수반되는 검색 알고리즘도 함께 연구되어야 할 것이다.

참고문헌

1. Kai-Fu Lee, Automatic Speech Recognition, Kluwer Academic Publishers, 1989

2. Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, "JANUS-III:Speech-To-Speech Translation in Multiple Languages", ICASSP-97, 1997.

3. 이 강성, "DMS 기법을 이용한 화자 독립 단독어 음성 인식," 광운대학교 신기술 연구소 논문집, Vol 26, 1997.

4. H.Ney, R.Haeb-Umbach, B.H.Tran, M.Oerder, "Improvements in Beam Search for 1000-Word Continuous Speech Recognition," ppl-9~12, ICASSP-92, 1992.

5. 문교부, 국어 어문 규정집, 대한 교과서 주식회사, 1997

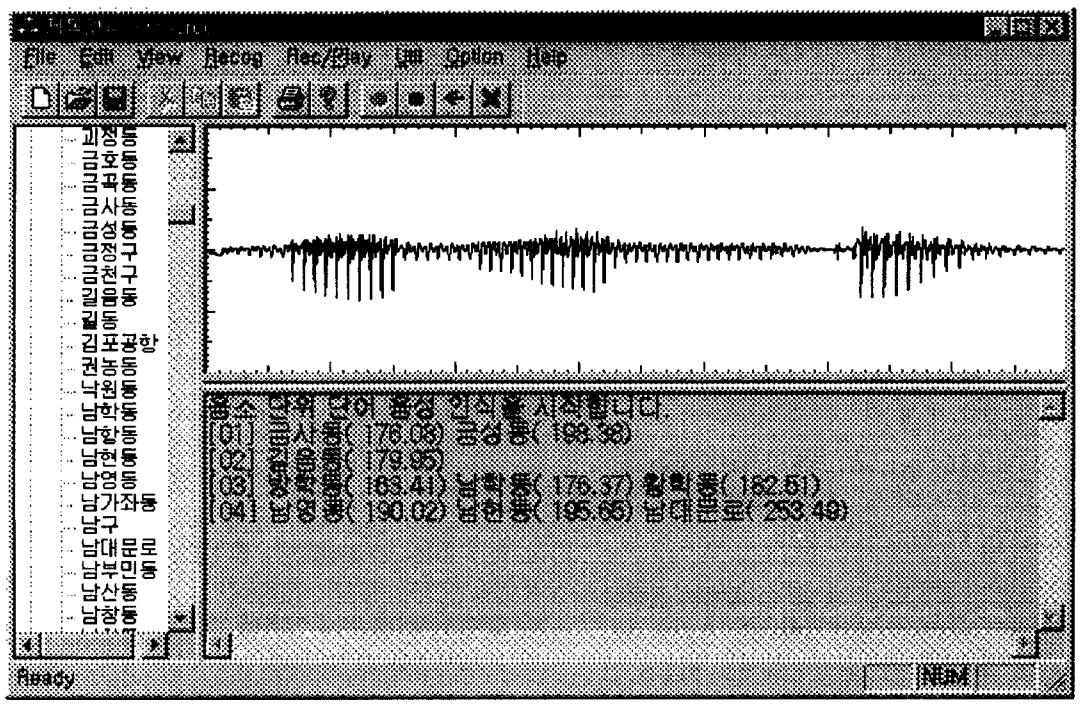

그림 4. 음성인식 프로그램 실행 예