무제한 단어 음성인식을 위한 모음열 사전의 구축

김동환, 윤재선, 홍광석
성균관대학교 전기전자 및 컴퓨터공학과 휴먼컴퓨터 연구실

A construction of vowel string dictionary for unlimited word speech recognition

Dong-Hwan Kim, Jeh-Seon Youn, Kwang-Seok Hong
HCI Lab, Electrical & Computer Engineering, SungKyunKwan University

kdh2k@popsmail.com, sunhci@ece.skku.ac.kr, kshong@yurim.skku.ac.kr

요약

기존의 제한적 단어 인식과는 달리 무제한 단어 음성인식에 있어서는 방대한 용량의 단어 모델을 참조로 인식이 이루어지게 되어, 참조모델과 입력패턴과의 비교를 위한 탐색시간이 너무 길어지게 된다. 본 논문에서 제한하는 방법은 무제한 단어 음성인식 시스템을 구축하기 위해 선행되어야 하는 모음열 사전을 구축하는 것이다. 음성인식시 입력패턴과 참조모델에 속한 모든 단어와의 비교를 수행하지 않고, 입력패턴의 모음열을 인식한 후, 인식된 모음열 단어들만을 참조모델에서 인식 후보로 두어 인식을 수행하게 하여 시간적인 측면에서의 효율성을 기하는 것이다. 결과적으로 본 연구 방법은 무제한 단어 음성인식에서의 실시간 처리라는 점에 주 목적을 두었다.

I. 서론

음성 처리 기술은 hardware의 발전과 더불어 급속한 성장을 이루고 있다. 이에 따라 기존의 제한적이었던 단어 수에서 이제는 어떠한 단어라도 인식해 낼 수 있는 무제한 단어 인식으로의 전환이 필요한 때이다. 보통 음성인식을 하기 위해 Template matching이나 HMM등에서는 모든 참조 단어와 입력 단어와의 비교가 필요하다. 그러나 참조 단어의 어휘 수가 대용량, 즉 1000단어 이상일 경우에는 입력 음성과 참조 모델간의 비교를 위한 많은 Memory와 Computation cost가 필요하게 되므로, 고성능의 컴퓨터가 아닌 이상 음성 인식의 실시간 처리에 어려움을 겪게 된다. 따라서 적은 계산량으로 얻을 수 있는 정보를 이용해야 할 것이다. 이에 본 논문에서는 음성인식의 실시간 처리를 위하여 후보 단어의 수를 줄이는 방법을 택하였다.

인식 대상의 어휘수가 적은 단어 인식에서는 기본 인식 단위로 주로 단어나 음절을 이용한다. 그러나 무제한 단어 인식일 경우 모든 단어에 대해 위의 인식 단위를 사용하는 것은 시간적인 측면에서의 효율성을 떨어뜨리는 결과를 낳게 된다. 따라서 좀 더 효율적인 방법으로 sub-word 단위를 사용하게 된다. sub-word 단위의 종류에는 allophone, phoneme, diphone, syllable 등이 있다. 이들의 장단점은 표 1.1과 같다.

위의 인식단위 중 본 논문에서 사용한 방법은 음절내에서 모음열을 추출하여 사용하는 것이다. 즉, 음성인식시 검출이 유리한 안정된 모음 구간을 찾아내고, 검출된 모음열에 해당하는 단어들만을 참조 모델과 비교하는 것이다. 따라서 전체 단어와의 비교시보다 상당량의 비교 과정을 줄일 수 있는 장점을 얻을 수 있다.

본 논문은 위의 수행을 가능하게 하기 위하여 우선적으로 구축되어야 하는 모음열 사전 구성 방법에 대해 논할 것이다.

표 1.1 음성 인식 단위의 특징

인식 단위	장점	단점
allophone (異音)	일부분은 음향학적으로 쉽게 구분	분류가 정확하지 않음 종류가 많음
phoneme (音素)	단어를 쉽게 phoneme으로 표시 가능 종류가 많음	음향학적으로 쉽게 분류 안됨 많은 음운 규칙 필요
diphone (二重單音)	transition정보 포함	음운 규칙의 적용 어려움 종류가 다양
syllable (音節)	쉽게 음절 위치를 찾아냄	정확한 음절 경계를 찾기 어려움 종류가 많음

II. 모음열 사전 구축 시스템

모음열 사전 구축 시스템의 전체적인 흐름을 그림 2.1에 나타내었다. 각 단계는 다음과 같다.

2.1 문서내 단어 추출

현재 open된 문서내의 모든 단어들을 공백을 기준으로 모두 추출해 낸다. 여기서 문서는 완성형을 기준으로 한다.

2.2 전처리 과정

추출된 모든 단어에 대해 다음과 같은 조건을 단계별로 적용시켜 불필요한 문자는 제거시킨다.

단어에 문장 부호, 숫자, 영문자, 특수문자, 한자 등이 포함되었을 경우에는 해당 문자만 제외시키고 단어만 추출해 낸다.

예) "갑니다!" → 갑니다　　國家 → (제외됨)
　　삽니.다 → 삽니다　　Korea한국 → 한국

본 시스템에서는 단어 길이를 2음절 이상 20음절 이하에 해당하는 단어만을 취급하였다. 그러나 실제로 10음절 이상의 단어들은 띄어쓰기의

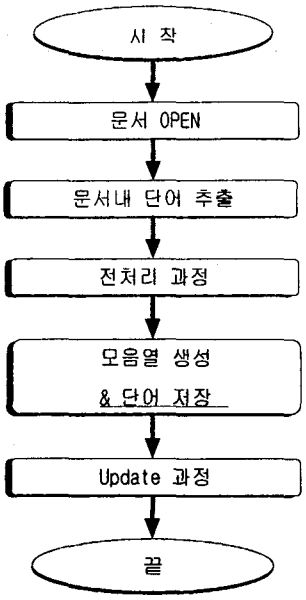

그림 2.1 모음열 사전 구축 과정

오류일 가능성이 많았다.

다음은 한글 단어에 적합한지를 검사한다. 먼저 각 단어의 음절별로 완성형 문자를 유니코드 문자로 변환시킨다. 각 음절의 음소(초성, 중성, 종성)가 모두 한글 유니코드의 범위(초성:0~18, 중성:0~20, 종성:0~27)내에 있어야 하며 그 외의 경우에는 제외시킨다.

예) ㅎ ㄹ ㅊ → 중성 없음 (제외됨)
　　ㅐ ㅛ ㅠ → 초성 없음 (제외됨)

2.3 모음열 생성과 단어 저장

유니코드로 변환된 단어의 각 음절의 모음에 해당하는 중성코드('ㅏ ㅐ ㅑ…')를 조사하여 초성코드에 해당하는 'ㅇ'과 결합시켜 모음열을 생성한다.

예) 우리집 → 우이이　　따뜻하다 → 아으아아

다음으로 모음열에 해당하는 단어를 파일에 추가 저장한다. 먼저, 단어의 음절길이를 측정한후 단어의 가장 첫 모음을 판별한다. 그런 후 모음

열 파일을 생성하고 해당 단어를 파일에 추가시킨다. 이 과정에서 각 단계별로 디렉토리가 생성되어진다. 이에 해당하는 과정을 그림 2.2에 나타내었다. 그림은 문서내에서 전처리 과정을 거쳐 추출된 단어들 중에 '타인의'와 '명예를'이라는 단어를 저장하는 과정을 보인것이다. '타인의'라는 단어는 3음절에 속하며 첫 모음이 '아'에 해당한다. 또한 모음열 생성과정을 통해 '아이의'라는 모음열 파일을 만들어낸다. 여기에 해당 단어 '타인의'를 추가시킨다. 이때 단어가 중복될 경우에는 해당 단어 수를 증가시키며 모든 단어들은 단어 파일에 따로 저장된다.

2.4 Update 과정

Update과정에서는 이전에 저장된 파일들의 해당 정보를 count 하게 된다. 총단어수와 중복되지 않는 단어수를 저장하게 되며, 각 음절수(2음절, 3음절, …)에 해당하는 단어수, 각 모음열(아아, 아애, …)에 해당하는 단어수, 그리고 전체 단어에 대한 모음 21개(아, 애, 야, …)의 개수도 저장하게 하였다.

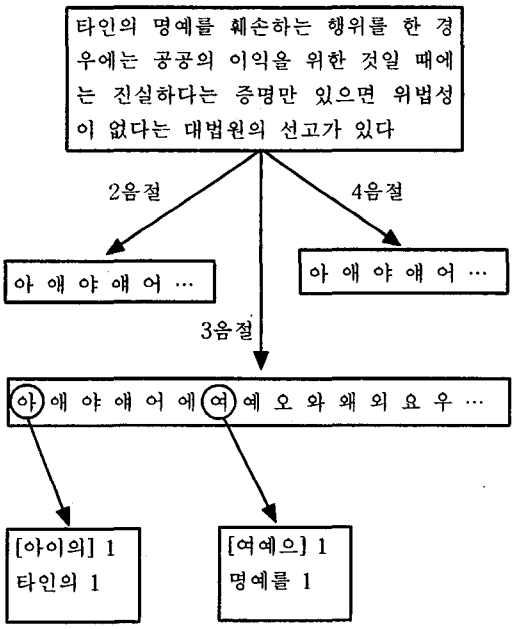

그림 2.2 모음열에 따른 단어 저장 과정

III. 모음열 탐색

모음열 탐색에 의한 단어 선택 과정을 그림 3.1에 나타내었다. 입력 음성으로부터 특징 파라메타가 추출된 후 모음 분류 시스템을 통해 모음열이 판단된다. 그 후에 Searching Algorithm을 통해 모음열 사전에서 후보 단어가 선택 되어진다. 예를 들어 모음열 '아아'에 대한 후보 단어로는 '가다', '바다', '잡다', '하다', '찾다'등이 올 수 있다. 이 후보 단어들에 음성 인식에서의 확률 알고리즘이 적용되어 결과적으로 인식 단어를 선택하게 되는 것이다. 여기서 탐색 알고리즘은 그림 2.2에서 보인 절차를 참고로 수행된다.

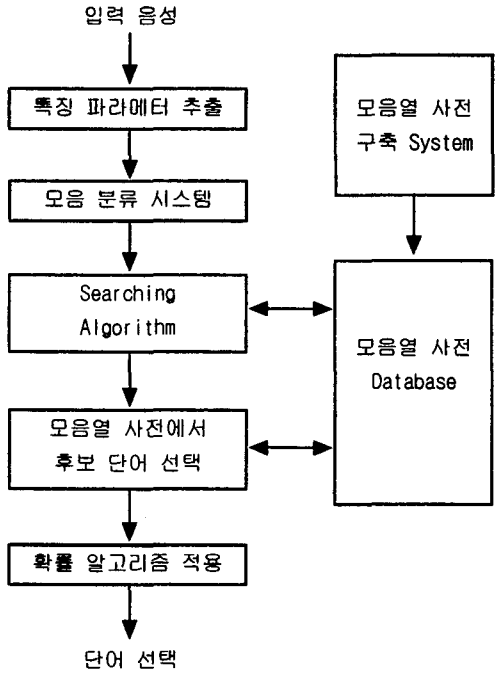

그림 3.1 모음열 탐색에 의한 단어 선택 과정

IV. 실험 및 결과

4.1 모음열 단어 추출

본 논문에서는 인터넷과 통신을 통해 수집된 약 50MB분량(소설, 수필, 대본, 성서, 일반 글모음 등)의 문서를 가지고 모음열 사전 구축 프로

그램을 수행시켰다. 그 결과 추출된 총단어수는 약 3백3십만개이고, 그에 따른 중복되지 않는 단어수는 약 5십3만개 정도였다. 그림 4.1은 총단어수 10만개당 중복되지 않는 단어수의 증가 형태를 나타낸 것이다. 문서 종류가 다양했던 만큼 단어수는 계속 증가하는 형태를 띄었다.

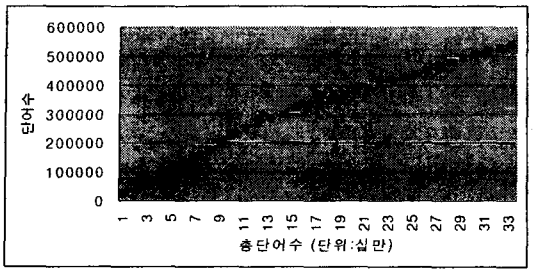

그림 4.1 총단어수와 단어수의 증가형태

2음절 이상 13음절 이하에 해당하는 단어수를 표 4.1에 나타내었다. 2음절, 3음절, 4음절에서 많은 수의 단어 분포를 보였고, 10음절 이상일 경우에는 띄어쓰기의 오류가 많았다.

표 4.1 음절 수에 따른 단어 분포

음절수	단어수(%)	음절수	단어수(%)
2음절	1297500 (39.25)	8음절	4302 (0.13)
3음절	1198949 (36.27)	9음절	1410 (0.04)
4음절	565564 (17.11)	10음절	546 (0.02)
5음절	192219 (5.81)	11음절	228 (0.01)
6음절	51590 (1.56)	12음절	140 (0.00)
7음절	14057 (0.43)	13음절	61 (0.00)

각 모음의 분포는 '아' 2백2십만개, '이' 1백5십만개, '으' 1백2십만개, '어' 1백만개, '오' 9만개, '우' 6만개의 순으로 측정되었다.

4.2 인식 실험

추출된 모음열 가운데 3음절에 해당하는 '아아이'계열의 단어들(예: '사람이', '갑자기'등)에 대해 인식 실험을 해 보았다. 이 단어들 가운데 출현 빈도수가 50이상인 단어 30개를 선정한 후 각각의 단어에 대해 확률 알고리즘을 적용하여 인식 성능을 측정하였다. 인식 결과 1후보, 2후보, 3후보에 대한 인식률이 표 4.2와 같이 측정되었다.

표 4.2 인식 성능

	1회	2회	3회	4회	5회	평균
1후보(%)	73	73	87	73	73	75.8
2후보(%)	80	90	90	90	83	86.6
3후보(%)	90	93	90	90	93	91.2

V. 결론

본 논문에서는 무제한 단어 음성 인식의 실시간 처리 기능을 향상시키기 위한 모음열 사전의 구축에 대하여 논하였다. 즉, 어떠한 입력 음성이 들어왔을 때 모음열을 판별하여 미리 구축해 놓은 모음열 사전에서 그에 해당하는 단어들만을 검색하고 비교하여, 전체 단어와 비교했을 때보다 시간적 · 계산적 측면에서 효율성을 높이는 것이다. 비록 본 논문에서는 모음열 사전을 구축하고 그에 해당하는 단어들로 인식 실험한 것에 그쳤지만, 모음열 탐색 부분과 형태소 분석과 같은 자연 언어 처리 부분이 인식 시스템에 추가 구현된다면 더욱 향상된 결과를 얻을 것이다.

참고문헌

[1] 윤재선, 정광우, 홍광석, "무제한 단어인식 시스템을 위한 VCCV분할에 관한 연구," 한국음향학회지 제19권 제1호, pp.103-106, 2000.

[2] 정용주, "대용량 단어 인식에서의 모음 분류를 이용한 시간 감축에 대한 연구," 한국과학기술원, 석사학위논문, 1990.

[3] 장동수, 서영훈, "음절에 기반한 한국어 형태소 분석기," 제5회 한글 및 한국어정보처리 학술발표논문집, 1993.

[4] 최기선외 5인, "한국어 철자 및 띄어쓰기 교정 시스템에 관한 연구 (II)," 과학 기술처, pp.23-35, 1992.