음절 복원 규칙과 형태소 분석을 이용한 음성인식 후처리

서상현†*, 김재홍†, 김해진†, 김미진†, 이상조†
† 경북대학교 컴퓨터공학과, † 경북대학교 국어국문학과

Post-Processing of Voice Recognition Using
Phonologic Rules and Morphologic analysis

Seo SangHyun†*, Kim JaeHong†, Kim HaeJin†, Kim MiJin†, Lee SangJo†
† Department of Computer Engineering, Kyungpook National University
† Department of Korean Language and Literature, Kyungpook National University

컴퓨터의 사용이 보편화됨에 따라 컴퓨터와 사용자 사이의 쉽고 자연스러운 의사 소통을 위한 자연어 인터페이스에 대한 연구가 활발히 진행되고 있다. 이 중에서 특히, 음성인식 분야는 음성명령, 받아쓰기 시스템 등 일반적인 컴퓨터 사용자의 요구를 충족시켜 줄 수 있는 분야로 주목을 받고 있다. 그러나 음성인식은 인식 자체 만으로는 인식률에 한계가 있으며, 인식 결과를 향상시키기 위해서는 후처리 단계가 필요하다.

본 논문에서는 음성 인식의 성능을 향상시키기 위해 음성 인식의 결과로 들어온 연속된 한국어 음성을 올바른 음절로 복원시켜 주는 시스템을 구현하였다. 이 시스템에서는 어절단위의 연속된 한국어 음성을 입력으로 받아 한국어 발음 규칙을 역으로 적용하여 원래의 음절로 복원시키고, 형태소 분석기를 이용하여 복원된 음절이 올바른지를 확인하고 수정한다. 초등학교 교과서에 나오는 문장을 대상으로 본 시스템의 성능을 실험한 결과, 90.42%의 복원율을 나타내었다. 현재 정확하게 복원이 되지 않는 것 중에는 동음이의어가 차지하는 비중이 크며, 이 문제는 구문분석이나 의미분석을 이용하여 어느 정도 개선할 수 있을 것으로 보인다.

1. 서 론

컴퓨터가 보편적으로 사용되기 시작하면서 컴퓨터와 사용자 사이의 통신을 더욱 간단하고 편리하게 하려는 노력이 계속되고 있다. 스캐너를 이용하여 대량의 문서를 입력할 수 있는 문자인식이나 사람의 음성을 직접 입력받을 수 있는 음성 인식 분야 등이 그 예이다.

그 중에서 음성 인식 기술은 단순한 음성을 통한 문서 입력 시스템뿐만 아니라, 상호회화에 의한 통역 전화, 강연의 동시통역, 국제전화교환 등의 음성 번역 시스템을 위해서도 필수적이다.

현재까지의 음성 언어 처리 연구는 음성 신호의 파형을 분석하여 단어 혹은 문장을 인식하는 인식 기술에 대한 연구가 주류를 이루고 있다. 그러나 이러한 음성 인식 기술만으로는 음성 언어를 문자 언어로 바꾸는데 한계가 있으며, 음성 인식기의 인식률을 높이는 데에도 한계가 있다. 이에 대한 기존의 연구로는 한국어 음성 언어에 대한 형태소 분석[1]을 들 수 있는데, 여기에서 제안된 시스템에서는 조음결합 현상을 처리하기 위해 음성 언어 사전을 구축하는 방법을 이용했으며, CYK 알고리즘을 이용해 형태소 분석을 시도했다. 그러나 이러한 방법은 모든 가능한 조음 결합현상을 사전에 등록해야 하는 어려움이 있으며, 따라서 어느 특정한 분야에 한정시키는 경우가 대부분이다. 그러므로 이러한 방법으로는 보편적으로 사용될 수 있는 음성 인식기나 음성 번역 시스템에 응용하기가 어렵다.

지금까지 음성인식 기술은 수천개의 단어를 인식하여 명령을 처리하는 수준의 상용화가 이루어져 있다. 하지만 음성을 이용한 문서 입력을 하기 위해서는 우선 연속된 음성을 올바른 음절로 복원시켜 주고 어절 단위로 띄어 주어야 한다. 이때 한국어에서 사용되는 음운 규칙 등의 국어학적 지식[2]과 자연어 처리 기술을 적용하면 처리율을 높일 수 있다.

본 논문에서는 음성인식기의 결과로 나온 연속된 음성 문자열을 가지고 자연어 처리에서 사용되는 형

태소 분석 기법과 발음할 때 발생되는 음운 규칙을 역추적하는 방법으로 음성 복원기를 구현하였다. 그리고 음성 문자열에 음운 규칙을 적용하여 복원된 결과가 올바른지를 판단하기 위해서 양방향 최장 일치에 의한 형태소 분석을 이용하였으며[3], 음운 규칙을 적용하는 과정에서 음운 규칙의 예외인 어절들은 예외 사전에서 해결하였다.

본 논문의 구성은 2장에서 전체 시스템의 구성에 대해서 알아 보고, 3장에서는 음절 복원 규칙과 그 적용의 예를 보여 준다. 4장에서 제안된 시스템을 실험한 결과를 살펴 보고, 마지막으로 5장에서는 결론 및 향후 연구 과제에 대해서 논한다.

2. 시스템의 구성 요소

본 논문에서 제안하는 한국어 음절 복원기는 한글을 발음할 때 발생하는 음운 규칙을 역으로 적용하고 그 결과를 가지고 복원기의 성능을 향상시키기 위해 형태소 분석을 행하였다. 제안된 복원기의 전체 구성도는 그림 1과 같다. 연속된 음성을 입력받아 한국어 음운 규칙을 역적용하기 위해서는 조사·어미와 어간의 구분이 필요하므로 이를 위하여 조사·어미사전과 선어말 어미 사전을 구성하였다. 또한 음운 규칙으로 처리할 수 없는 단어들을 위해 예외사전을 구성하였으며, 음운 규칙을 이용하여 올바른 음절을 복원하는 과정에서 생성 가능한 후보들의 수를 줄이기 위해 음절 사전을 구성하였다. 음절 사전은 음운을 가지고 조합가능한 모든 음절 중에서 한국어에서 사용되고 있는 음절만을 모은 것이다.

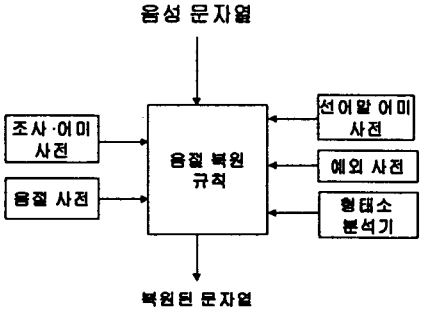


그림 1. 전체 시스템 구성도

2.1 음성 문자열

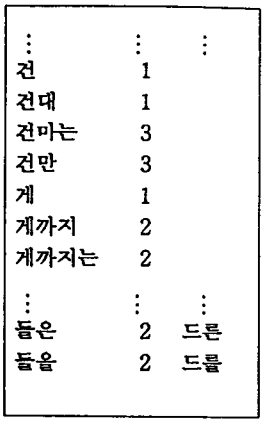
음절 복원기의 입력은 "나는 바블 먹꼬 학교에 간따"와 같은 음성 문자열이다. 이 음성 문자열은 음성 인식기에서 음성 그대로를 인식한 결과이다. 그러나 연속된 음성 문자열 자체로는 정보로서의 가치가 없기 때문에 이 음성 문자열을 올바른 음절로 구성된 문장으로 변환시켜 주어야 한다. 그리고 입력 단위는 어절로 하였는데, 입력을 언절 단위(말을 하는 단위)로 할 경우 말하는 사람이나 속도에 따라서 달라지므로 일관성이 없어진다. 따라서 본 논문에서는 어절 단위의 연속된 음성 문자열을 실험 대상으로 하였다.

본 논문의 입력인 음성 문자열은 초등학교 교과서 문장을 음운 규칙을 적용하여 음성문자열나는 대로 적은 음성 문자열로서 고저나 장단은 고려하지 않았다.

2.2 조사·어미 사전

한국어는 체언이나 용언에 조사나 어미가 결합하여 어절을 이루는 첨가어이기 때문에 이러한 실질 형태소와 형식 형태소간의 결합에서 음운 변화가 빈번히 발생한다. 따라서 연속된 한국어 음성 문자열에서 용언과 어미, 체언과 조사를 분리한다면 음운 규칙을 역으로 적용하기가 쉽다.
따라서 본 논문에서 이러한 실질 형태소와 형식 형태소를 분리하기 위해 조사·어미 사전과 선어말 어미 사전을 사용하였다.

조사·어미 사전은 그림 2에서 보는 바와 같이 조사·어미 문자열, 조사·어미 유형 정보 그리고 조사·어미의 음성 문자열로 구성되어 있다. 조사·어미 문자열은 한국어에서 사용되고 있는 조사나 어미들을 수록한 것으로 두 개 이상 결합된 조사나 어미도 포함되어 있다. 조사·어미 유형 정보는 조사·어미 문자열이 조사로 사용되는 것인지, 어미로 사용되는 것인지 또는 조사·어미 둘다로 사용될 수 있는 것인지에 대한 정보이다. 1은 조사를 의미하고, 2는 어미를 나타내며, 3은 조사 또는 어미로 사용될 수 있음을 보여준다. 마지막으로, 조사·어미의 음성 문자열은 조사·어미 문자열을 음성문자열나는 대로 적은 것으로 두 문자열이 같을 경우에는 생략하였다. 조사나 어미를 음운규칙을 적용해 가면서 찾는 방법도 있지만, 단일 조사·어미내에서나 조사·어미가 두 개 이상 결합할 때 연음이나 경음화 등의 음운 현상이 일어나서 복잡한 처리 과정을 거쳐야 하므로 본 논문에서는 조사·어미의 음성 문자열을 사전에 수록하여 처리 과정을 간단히 하였다.

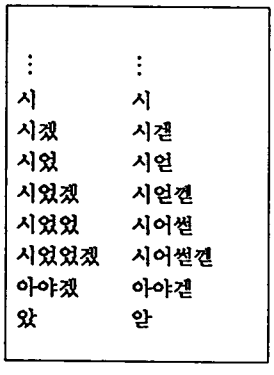


⋮	⋮	⋮
건	1	
건대	1	
건마는	3	
건만	3	
게	1	
게까지	2	
게까지는	2	
⋮	⋮	⋮
들은	2	드른
들을	2	드를

그림 2. 조사·어미 사전의 예

2.3 선어말 어미 사전

한국어에서 선어말 어미는 어미 앞에 붙어 시제나 높임 등을 나타내는 것으로, 어미와 용언의 어간을 구분해 준다. 선어말 어미 사전은 그림3에서 보는 바와 같이 선어말 어미 문자열과 선어말 어미 음성 문자열로 구성된다.



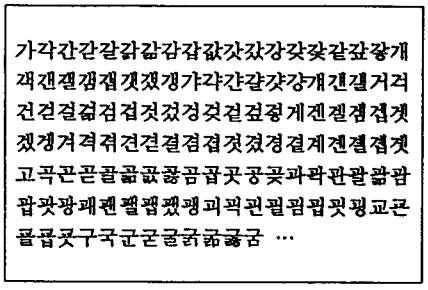
⋮	⋮
시	시
시겠	시겯
시었	시얻
시었겠	시얻겯
시었었	시어썯
시었었겠	시어썯겯
아야겠	아야겯
았	알

그림 3. 선어말 사전의 예

2.4 음절 사전

음절 사전은 음운을 가지고 조합가능한 모든 음절 중에서 한국어에서 사용되고 있는 음절만을 모은 것으로, 상용조합 2350자로 구성되어 있다. 음운 규칙의 역적용 과정에서 끝음성문자열 현상이나 연음 규칙으로 종성을 복원할 때 여러 가지 가능한 후보들이 생성될 수 있는데 이 후보들 중에는 사용되지 않는 글자들도 있다. 따라서 음절 사전을 이용하여 불필요한 음절 후보들을 제거할 수 있다.

그림 4는 음절 사전의 예이다.



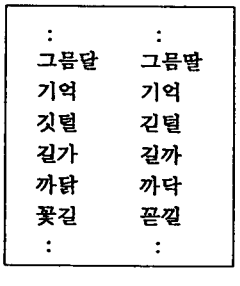
가각간갇갈갉갊감갑값갓갔강갖갗같갚갛개
객갠갤갬갭갯갰갱갸갹갼걀걋걍걔걘걜거걱
건걷걸걺검겁것겄겅겆겉겊겋게겐겔겜겝겟
겠겡겨격겪견걷결겸겹겻겼경곁계곈곌곔곗
고곡곤곧골곪곬곰곱곳공곶과곽관괄괆괌
괍괏광괘괜괠괩괬괭괴괵괸괼굄굅굇굉교굔
굘굡굣구국군굳굴굵굶굻굼 …

그림 4. 음절 사전의 예

2.5 예외 사전

예외 사전은 조사·어미 사전, 선어말 어미 사전, 음절 사전을 가지고 음절 복원 규칙을 적용하여 올바른 음절로 복원시킬 수 없는 경우를 위해 구성한 것이다. 음절 복원 규칙을 이용하여 올바른 음절로 복원할 수 없는 것들로는, 첫 번째로 음절 복원 규칙의 예외가 되는 경우인 "글쎄", "엄마" 등을 들 수 있다. "글쎄", "엄마"와 같은 경우는 음절 복원 규칙이 적용될 경우 "글쎄"는 경음화 현상을 복원하는 과정에서 "글세"로, "엄마"는 자음동화를 복원하는 과정에서 "업마"로 된다. 이러한 음성 문자열은 음절 복원 규칙이 적용되는 것을 막아야 하므로 "글쎄"와 "엄마"를 예외 사전에 두어 음절 복원 규칙보다 먼저 처리한다. 두번째로 "담넉"과 같이 음절 복원 규칙의 적용 순서로 인해 잘못된 음절로 복원되는 경우가 있다. "담넉"의 경우 자음동화 현상을 처리하면서 'ㄴ'이 'ㅁ'뒤에서 'ㄹ'로 바뀐다는 규칙을 이용하여 복원하여야 하지만, 'ㅁ'이 'ㄴ'이나 'ㅁ'앞에서 'ㅂ'으로 바뀐다는 규칙이 먼저 적용되었기 때문에 "담넉"으로 잘못된 결과를 도출시킨다. 따라서, 이러한 음성 문자열들도 예외 사전을 이용하여 처리한다.

예외 사전은 그림 5와 같이 문자열과 그것의 음성 문자열로 구성된다.



:	:
그믐달	그믐딸
기억	기억
깃털	긷털
길가	길까
까닭	까닥
꽃길	꼳낄
:	:

그림 5. 예외 사전의 예

2.6 형태소 분석기

형태소 분석기는 음절 복원 규칙을 적용하는 과정과 복원 규칙을 모두 적용하고 난 뒤 결과를 확인하기 위해서 사용된다. 본 논문에서는 양방향 최장 일치법을 이용한 형태소 분석기를 이용하였다[3].

3. 음절 복원기

3.1 음절 복원 규칙 및 적용 순서

한국어에서 많은 단어들은 발음될 때 하나의 음운 규칙뿐만 아니라 여러 개의 음운 규칙들이 순서에 따라 적용될 수 있다. 본 논문은 음성 문자열로부터 음절을 복원하는 것이므로 이를 위해서는 음운 규칙을 역적용하여야 한다. 따라서, 본 논문에서 음절 복원을 위해 이용하는 음운 규칙의 역적용을 음절 복원 규칙이라고 하였다.

한국어 음성 문자열에 대해서 음절 복원 규칙을 적용하기 위해서는 우선 어떤 순서에 의해서 음절 복원 규칙이 적용되는지를 알아야 한다. 왜냐하면 음절 복원 규칙이 적용되는 순서가 바뀌면 잘못된 결과가 도출되기 때문이다.

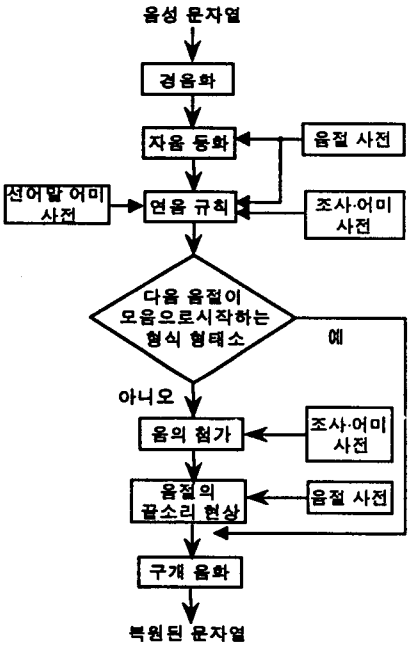


그림 6. 음운 복원 순서도

음절 복원 규칙의 적용순서는 위의 그림 6과 같다. 이때 한국어의 음운 현상으로 존재하더라도 사용 빈도가 낮은 경우는 예외사전으로 구성하는 것이 더욱 효율적이다.

음성 문자열을 입력받아 가장 먼저 처리해야 할 음절 복원 규칙은 경음화 현상에 대한 역적용이다. 경음화는 특정 음소들간의 결합에 따른 음운 현상이므로 음소들의 배열이 규칙에 해당되는지를 살펴 처리한다.

자음 동화는 음절 복원 규칙을 적용하는 과정에서 음절 사전을 이용한다. 즉 자음 동화를 역적용한 후의 음절이 한글에서 사용되지 않는 음절일 경우에는 잘못 복원된 음절로 보고 그 규칙과 관련된 다른 음소로 바꾼 후 다시 음절 사전을 검색한다. 사용되고 있는 글자일 경우에는 복원될 음소간에 우선 순위를 두어 처리한다. 그러나 자음 동화를 역적용할 때 "선생님", "생물"과 같이 규칙으로 해결할 수 없는 경우도 있다. 이러한 경우는 음소 간의 정보나 품사 정보만을 가지고는 해결할 수 없으므로 예외 사전을 구성하여 처리하였다. 그리고 "신라"나 "물난리"와 같이 'ㄴ'은 'ㄹ'의 앞이나 뒤에서 'ㄹ'로 음성문자열가 난다는 규칙은 앞음절 종성의 'ㄹ'이 바뀌었는지 뒷음절의 초성의 'ㄹ'이 바뀌었는지 알 수가 없다. 이 경우에는 가능한 후보들을 형태소 분석하여 성공하는 후보를 선택한다. 위의 예에서 "실라"는 "신라"와 "실나"로 복원될 수 있는데 이 때 형태소 분석에 성공한 "신라"를 올바른 복원으로 본다.

연음 규칙은 체언이나 용언이 모음으로 시작하는 조사나 어미와 결합할 때 발생하는 음운 현상으로 조사나 어미를 구별해야 하는데 한국어에서는 조사가 생략되기 쉬우므로 처리가 어려워진다. 그러나 "개구리가"와 같이 "개굴"+"이가"나 "개구리"+"가"로 분석되는 경우에는 조사 정보만으로 처리가 어려우므로 두 후보를 형태소 분석하여 성공한 후보를 선택한다. 분석에 실패한 후보는 분석하기 전으로 되돌리거나 다른 조사·어미로 분리를 한다.

음의 첨가는 합성어 및 파생어에서, 앞 단어나 접두사가 자음으로 끝나고 뒷 단어나 접미사의 첫 음절이 '이, 야, 여, 요, 유'인 경우에 'ㄴ'음을 첨가하여 발음하는 규칙이다[4]. 입력된 음성 문자열 가운데 '니, 냐, 녀, 뇨, 뉴'가 있으면 그 뒷음절과 함께 조사·어미 사전을 탐색하여 실패할 경우 'ㄴ'이 첨가된 것으로 처리한다.

음절의 끝음성문자열 현상은 한국어에서 받침으로 나는 음성문자열가 'ㄱ,ㄴ,ㄷ,ㄹ,ㅁ,ㅂ,ㅇ'의 7개 자음으로 한정된다는 것이므로[4], 그 외의 자음들은 7개의 음 중에서 하나로 음성문자열가 나게 된다. "갇",

"갓", "갔", "갖", "갗", "갛"이 모두 "간"으로 음성문자열나기 때문에 음절 정보만을 가지고는 이를 해결하기가 어렵다. 따라서 음절의 끝음성문자열 현상을 역적용할 때 조사·어미 유형정보와 음절 사전을 이용한다. 또 용언의 어간에는 나타나지만 체언에는 나타나지 않는 음절 정보를 이용하고 한국어의 종성에서 나타날 확률이 높은 자음에 우선순위를 줌으로써 처리한다.

3.2 규칙 적용의 예

본 시스템에서 다음과 같은 음성 문자열을 입력받았을 경우의 처리 과정을 살펴 보면 다음과 같다.

음성 문자열 "장소의 바뀜를 생가카며 미영이가 지베서 출발하여 도라을 때까지의 이를 자세히 말하여 봅씨다"가 입력되었을 경우 여기에 예외 규칙이 첫 번째로 적용되어 "장소의 바뀜를 생가카며 미영이가 지베서 출발하여 돌아을 때까지의 이를 자세히 말하여 봅씨다"가 되어서 그 다음 단계로 넘어간다. 그 다음으로는 경음화의 역적용이 이루어져야 하며 그 결과는 "장소의 바뀜를 생가카며 미영이가 지베서 출발하여 돌아을 때까지의 이를 자세히 말하여 봅시다"가 된다. 그 다음 단계인 자음 동화의 음절 복원 규칙은 그 적용예가 없으므로 다음의 연음 규칙의 역적용을 받게 되고 그 결과 "장소의 바뀜을 생각하며 미영이가 집에서 출발하여 돌아을 때까지의 일을 자세히 말하여 봅시다"가 되며 그 이후에는 순서도 상에 나타나는 다른 복원 규칙들이 적용되는 어절이 없으므로 복원된 문자열은 다음과 같다.

"장소의 바뀜을 생각하며 미영이가 집에서 출발하여 돌아을 때까지의 일을 자세히 말하여 봅시다"

4. 실험 및 결과

위 시스템은 DEC 3000 WorkStation에서 c언어로 구현되었고, 초등학교 교과서 1만 어절을 대상으로 하였다. 실험과정은 음운 규칙을 적용하여 음성문자열나는 대로 고쳐 적은 음성 문자열을 본 논문에서 제안한 음절 복원기로 복원시키고 이를 원래의 문장과 비교한 것이다. 실험 결과는 표 1과 같다.

표 1. 실험 결과

	복원율
음절 단위	94.30%
어절 단위	90.42%

음절 단위의 복원율은 94.30%로 나타났고 어절 단위는 이보다 낮은 90.42%의 복원율을 나타냈다.

본 시스템으로 정확하게 복원시킬 수 없는 것 중에는 동음이의어가 차지하는 비율이 크며 이러한 동음이의이는 구문분석이나 의미분석을 하지 않으면 정확하게 복원시킬 수 없는 것이다.

5. 결 론

본 논문에서는 음성 인식의 후처리로서, 연속된 음성 문자열이 입력되었을 때 올바른 한국어 음절로 복원시켜 주는 시스템을 제안하였다. 제안한 음절 복원기는 음절 복원 규칙과 조사·어미 사전, 선어말 어미 사전, 예외 사전, 음절 사전을 이용하여 음절을 복원하고 형태소 분석기를 통해 복원된 음절의 정확성 여부를 판단하여 재수정한 것이다.

초등학교 교과서 1만 어절을 대상으로 실험한 결과 90.42%의 복원율을 나타내었다. 음절 복원기로 복원시키지 못한 것에는 동음이의어가 차지하는 비율이 큼을 볼 수 있는데, 구문분석과 의미분석을 이용하면 복원율을 향상시킬 수 있을 것으로 보인다. 그리고 사람의 말음성문자열에는 음성뿐만 아니라 고저나 장단도 가지고 있으므로 음성 인식을 할 때 이들을 고려하여 이용한다면 보다 높은 인식률을 가져올 수 있을 것이다.

본 논문에서 제안한 음절 복원기를 음성 인식 결과의 후처리로 이용하여 인식률을 높임으로써 단순한 음성을 통한 문서 입력 시스템뿐만 아니라 상호 회화에 의한 통역전화, 강연의 동시통역, 국제 전화 교환 등에 응용할 수 있다.

참 고 문 헌

[1] 김경희, "한국어 음성 언어 처리를 위한 음소단위 인식과 형태소 분석기의 결합", 포항공과대학교 대학원 석사 학위 논문, 1994

[2] 허웅, "국어 음운론" 샘문화사, 1985

[3] 최재혁, "양방향 최장일치법에 의한 한국어 형태소 분석기의 구현", 경북대학교 대학원 박사 학위 논문, 1993

[4] 최기호, "새 한글맞춤법 길라잡이", 토담출판사, 1994