한국어 음가의 표기 복원을 위한 표기 후보 생성 및 감소에 관한 연구

이 상 범[†]·박 성 현[††]

요 약

음절 복원은 음성 인식 장치에서 인식된 음가열을 발성 이전의 표기 형태로 복원하는 과정이다. 본 논문에서는 음절 복원 과정을 위하여 표준 발음법을 기반으로 음절 복원 규칙을 작성하였다. 음절 복원 규칙을 이용하여 표기 후보 집합의 생성 방법을 연구하였다. 또한 생성된 표기 후보의 수를 감소시키기 위하여, 비표기 음절을 포함한 표기 후보 감소, 비어휘 음절을 포함한 표기 후보 감소, 비어간 음절을 포함한 표기 후보 감소의 3단계 감소 과정을 제안하였다. 제안된 방법을 통하여 실험한 결과 평균 74%의 표기 후보 감소율을 나타내었다.

A Study On Generation and Reduction of the Notation Candidate for the Notation Restoration of Korean Phonetic Value

Sang-Burm Rhee[†] · Sung-Hyun Park[††]

ABSTRACT

The syllable restoration is a process restoring a phonetic value recognized in a speech recognition device with the notation form that a vocalization is former. In this paper a syllable restoration rule was composed of a based on standard pronunciation for a syllable restoration process. A syllable restoring regulation was used, and a generation method of a notation candidate set was researched. Also, A study is held to reduce the number of created notation candidate. Three phases of reduction processes were suggested. Reduction of a notation candidate has the non-notation syllable, non-vocabulary syllable and non-stem syllable. As a result of experiment, an average of 74% notation candidate decrease rates were shown.

키워드 : 음절 복원(Syllable Restoration), 표기 복원(Orthographic Transcription), 복원 후보 축소(Reduce Syllable Restoration Candidate), 한글(Hangul), 음성 인식(Voice Recognition)

1. 서 론

일반적인 음성 인식 시스템은 전처리(preprocessing)과정, 특징 추출(feature extraction) 과정, 정합(matching) 및 후처리(postprocessing) 과정으로 이루어진다[1-3]. 음절 복원은 발성된 음가를 발성 이전의 표기로 복원하는 과정으로 음성 인식 시스템의 후처리에 사용되어 인식된 음가를 발성 이전의 표기로 변환하기 위한 방법으로 사용되며, 정합 부분에서 오인식된 부분을 교정하기 위하여 사용된다.

본 논문에서는 인식장치에서 인식된 음가열의 띄어쓰기 정보 복원[4-6]이 완료된 음가열에 대한 음절 복원에 관한 연구를 수행하였다. 인식된 음가열을 후처리 단계에서 발성 이전의 표기 형태로 변환하기 위하여, 표준 발음법을 모델링하여 생성된 표기-음가 변환표를 기반으로 음절 복원 규칙을 생성하고 이를 이용하여 표기 후보 집합을 생성하는 방법을 연구하였다. 또한 생성된 표기 후보 집합의 크기를 감소하기 위한 3가지 방안에 대하여 제안하였다. 1차 감소 방안은 표기 후보 집합 생성 후에 일상의 문자 언어 생활에서 사용되지 않는 완성형 한글 음절 이외의 음절을 포함한 표기 후보를 감소시키는 방안이다. 2차 감소 방안은 표기 후보가 생성된 후에 코퍼스(corpus)와 국어 사전을 기반으로 어절의 첫 번째 음절에 사용되지 않는 음절을 포함한 표기 후보와 어절 내에서 사용되지 않는 음절과 종성 초성쌍을 포함한 표기 후보를 삭제하여 표기 후보 집합의 크기를 감소시키는 방안이다. 3차 감소 방안은 형태소 분석 후에 추가되는 형태소 정보를 이용하여 어간에 나타나지 않은 음절을 포함한 표기 후보를 감소시키기 위한 방안이다.

본 논문의 연구 결과를 확인하기 위하여 제안된 음절 복원 규칙과 표기 후보 감소 방안에 따른 음절 복원 시스템을 구현하였다. 표기 음절이 연속적으로 발음될 때 일어나는 음운 변동이 반영된 음가열을 입력하면 제안된 방법을

※ 이 연구는 2003년도 단국대학교 대학연구비의 지원으로 연구되었음.
† 종신회원 : 단국대학교 전기전자컴퓨터공학부 교수
†† 준 회 원 : 단국대학교 대학원 전자컴퓨터공학과
논문접수 : 2003년 11월 28일, 심사완료 : 2004년 2월 13일

통하여 음절 복원이 이루어진다. 실험을 통하여 제안된 생성 음절 복원 규칙에 따라 발성전 표기를 포함한 표기 후보 집합이 생성됨을 보이고, 감소 방안에 따라 표기 후보의 수가 감소됨을 확인하였다.

2. 음절 복원 후보 생성

2.1 음절 복원 규칙의 생성

2.1.1 음절 복원에 관한 기존 연구

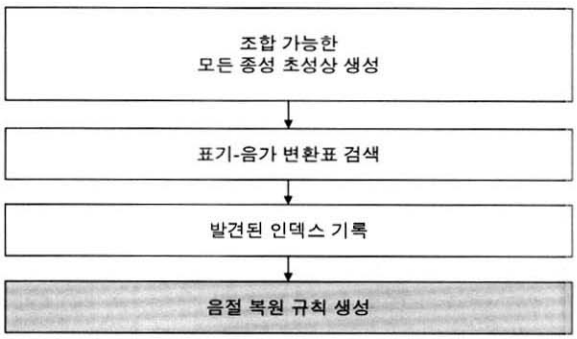
한국어가 연속적으로 발음될 때 여러 가지 음운 변동이 일어난다. 음절 복원은 이러한 음운 변동이 반영된 음가열을 변동 이전의 문자열로 다시 복원시켜 주는 과정이다[7].

음절 복원 규칙으로 읽기 규칙을 역으로 적용한 방식[8]은 한국어를 발음할 때 발생하는 음운 규칙을 역으로 적용하여 음절 복원을 수행한다. 복원을 위해 조사어미 사전, 음절 사전, 선어말 어미 사전, 예외 사전을 이용하였다. 이 방식은 규칙의 적용 순서가 복잡하여 알고리즘이 복잡하다는 단점이 있다[7]. 발음열 사전을 이용한 방식[9]은 인식 결과로 나온 음소열을 표제어로 하여 그 음소열이 발음될 수 있는 모든 단어를 사전에 수록하거나 발음 규칙을 기반으로 발음열 사전을 만들어 사용하였다. 그러므로 정해져 있는 도메인을 확장할 경우 사전을 구축하는 시간이 많이 걸리고 사전량이 커지는 단점이 있다. 자소 단위 사전을 이용하여 형태소 단계에서 음운 변동을 처리한 방식[10]은 일부 음운 변동을 규칙으로 정의하여 처리하였고, 자소 단위 사전 검색을 통해 형태소 분석 단계에서 복원 후보를 결정하는 방식이기 때문에 사전 검색 횟수가 많다는 단점이 있다[7].

2.1.2 제안된 연구 방법

본 논문에서는 음절 복원 규칙의 적용 알고리즘을 단순화하고 사전 검색 횟수를 줄이며 음소열 또는 발음열 사전을 사용하지 않아 도메인 확장에 유연한 대처가 가능하도록 하기 위하여 읽기 규칙을 역으로 적용한 방식을 개선하여 음가의 종성 초성 정보만으로 복원 규칙의 검색이 가능한 음절 복원 규칙을 생성한다

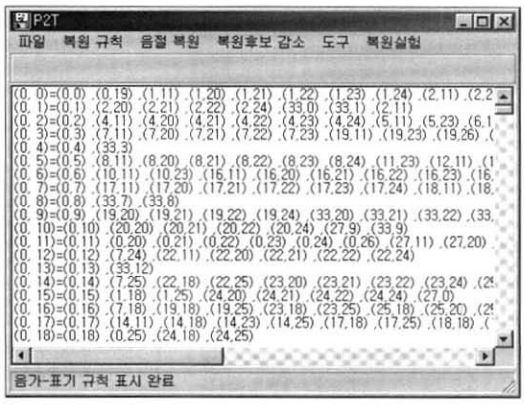
본 논문에서 사용한 표기-음가 변환표[11-13]는 표준 발음법을 분석하여 생성된 한국어의 표준 발음에 근거한 한국어 음성합성을 위한 읽기 규칙이다. 읽기 규칙을 역으로 적용하면 인식기에서 인식된 음가열에 대하여 발성 이전의 표기로 복원 할 수 있게 된다[7]. 음절 복원 규칙은 표기-음가 변환표를 역으로 적용하는 것으로 생성할 수 있다. 표기-음가 변환표는 앞음절의 종성과 뒷음절의 초성 정보와 형태소 정보를 인덱스로 하여 검색된 종성, 초성을 한번의 테이블 탐색으로 찾아 음가 변환을 완료하는 구조를 갖고 있다[11-13]. 음절 복원 규칙은 표기-음가 변환표에서 음가 변환된 종성-초성 쌍의 표기를 만들어내는 인덱스를 찾는 방법으로 음절 복원 규칙이 구성된다. 음절 복원 규칙을 생성하는 방법은 (그림 1)과 같다.



(그림 1) 음절 복원 규칙 생성 구조도

(그림 1)은 음절 복원 규칙 생성 과정을 나타내고 있다. 음절 복원 알고리즘은 표준 발음법을 분석하여 생성된 표기-음가 변환표를 바탕으로 한글 음절 순서에 따라 조합 가능한 모든 초성 종성 순서쌍을 만들어 낸다. 표기-음가 변환표에서 '종성, 초성' 쌍에 해당하는 인덱스를 검색한다. 검색된 인덱스가 초성 종성쌍의 발음을 생성하는 음절 복원 규칙이 된다.

(a) 한글 자모 나타낸 음절 복원 규칙



(b)유니코드로 나타낸 음절 복원 규칙

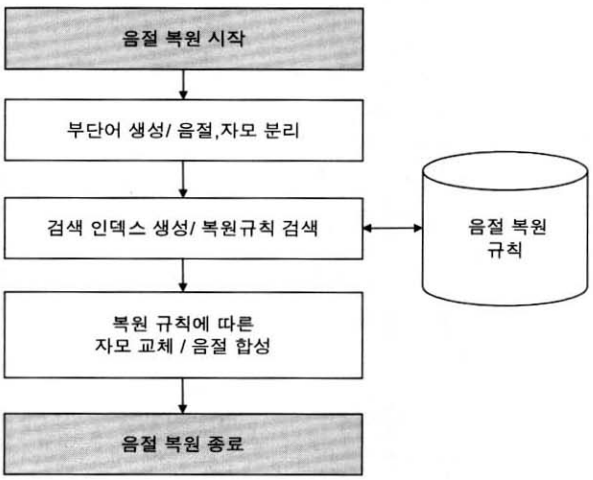
(그림 2) 음절 복원 규칙의 일부

(그림 2)는 음절 복원 규칙 생성과정을 통해 생성된 음절 복원 규칙의 일부를 나타낸 그림이다. 입력된 음가의 앞음절의 종성과 뒷음절의 초성의 유니코드 인덱스[14-15]로 해당하는 음절 복원 규칙을 찾아낼 수 있다

2.2 표기 후보 생성과 감소 방안

2.2.1 표기 후보의 생성

인식기를 통하여 인식된 음가열에서 음절 복원 규칙에 따라 표기 후보를 생성하는 과정은 (그림 3)과 같다.



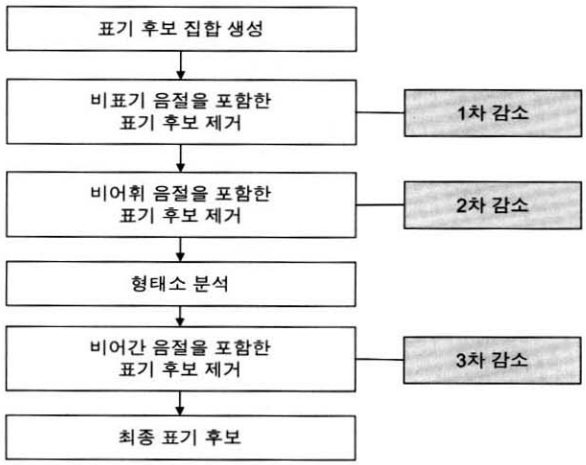
(그림 3) 표기 후보 생성 방법

(그림 3)은 표기 후보 생성의 전체 구성을 나타내고 있다. 부단어 생성 모듈에서는 인식된 음가열에서 띄어쓰기 단위로 어절과 음절 자모를 각각 분리해 낸다. 검색 인덱스 생성 모듈에서는 음절 단위로 분리된 자모에서 앞음절의 종성과 뒷음절의 초성을 하나의 쌍으로 묶어 검색 인덱스를 생성한다. 생성된 검색 인덱스로 음절 복원 규칙을 검색하여 해당 검색 인덱스에 해당하는 음절 복원 규칙을 찾아낸다. 찾아낸 음절 복원 규칙에 따라 자음과 모음을 교체한 후 음절로 합성하면 음절 복원 규칙에 따른 표기 후보가 생성된다.

"학꾜"를 예로 설명하면 음절 분리 모듈에서는 '학', '꾜'라는 2개의 음절로 분리하고, 자모분리 모듈에서는 '학-ㅎ, ㅏ, ㄱ' ,'꾜-ㄲ, ㅛ'의 자모로 분리해 낸다. 검색 인덱스 생성 모듈에서는 앞글자의 종성인 'ㄱ'과 뒷글자의 초성인'ㄲ'을 검색 인덱스인 (1, 1)의 검색 인덱스를 생성한다. 음절 복원 규칙 검색 모듈에서는 음절 복원 규칙을 검색하여 (1, 1)에 해당하는 "(1, 0), (1, 1), (2, 0), (2, 1), (3, 0), (3, 1), (9,0), (9, 1), (24, 0), (24, 1), (34, 1)"의 복원 규칙을 검색해 내고 이것은 "(ㄱ, ㄱ), (ㄱ, ㄲ), (ㄲ, ㄱ), (ㄲ, ㄲ), (, ㄱ), (, ㄲ), (ㄹ, ㄱ), (ㄹ, ㄲ), (ㅋ, ㄱ), (ㅋ, ㄲ), (어간받침ㄹ, ㄲ)"의 규칙을 갖는다. 음절 복원 규칙에 따라 자모를 대체하면 '학교', '학꾜', '핚교', '핚꾜', … 등의 표기 후보 집합이 생성된다. "학꾜"의 음절 복원 예를 4장 실험 및 고찰의 (그림 8)에 나타내었다.

2.2.2 표기 후보 감소 방안

형태소 분석과 구문 분석은 자연어 처리에서 빠져서는 안될 중요한 부분이지만 형태소 분석 과정에서는 입력된 분석 대상의 수에 대하여 2배 이상의 분석 결과를 갖게 되고, 이를 조합하여 생성된 문장 후보는 형태소 분석에서 생성된 분석 후보의 제곱에 해당하는 구문 분석 후보를 갖게 된다[16-17]. 생성된 문장 후보가 문법적으로 타당한지를 검색하는 구문 분석단계는 형태소 분석보다 훨씬 더 많은 시간을 소비하므로, 형태소 분석이전의 표기 후보 생성 과정에서 최소한의 분석 후보만이 생성되도록 하는 방안에 대한 연구가 필요하다.



(그림 4) 표기 후보 감소 과정

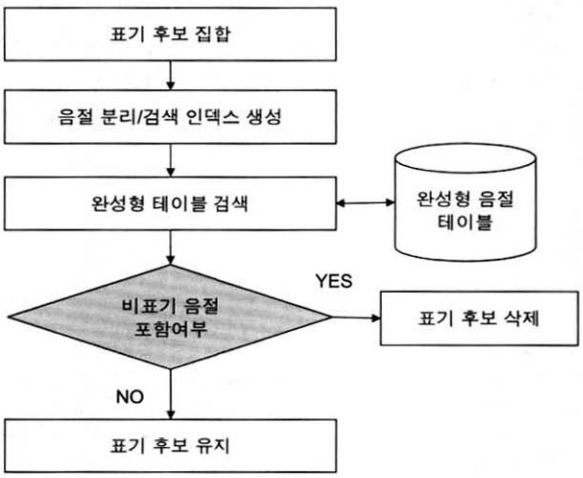
(그림 4)는 표기 후보 생성 과정에서 적용 가능한 감소 방안의 적용 시점을 나타내고 있다. 1차 감소 방안은 표기 후보 생성 후에 일상의 문자 언어 생활에서 사용되지 않는 완성형 한글 음절 이외의 음절을 포함한 표기 후보를 감소시키는 방안이다. 2차 감소 방안은 표기 후보 생성 후 감소 방안으로 표기 후보가 생성된 후 형태소 분석 이전에 표기 후보의 수를 감소하기 위하여 첫 번째 음절과 마지막 음절에 사용되지 않는 음절을 포함한 표기 후보와 어절 내에 사용되지 않는 음절 쌍을 포함한 표기 후보를 삭제하여 표기 후보의 수를 감소시키는 방안이다. 3차 감소 방안은 형태소 분석후에 어간에 사용되지 않는 음절을 포함한 표기 후보를 감소시키기 위한 방안이다.

3. 표기 후보 감소

3.1 비표기 음절 감소

표기 후보 생성 과정에서 보인 예와 같이 생성된 표기

후보 중에서는 실생활의 한글 표기에서는 타나나지 않은 음절을 포함한 표기 후보들이 생성된다. 본 논문에서는 일상의 언어 생활에서 나타나지 않은 문자 즉, 완성형 한글에서 사용되는 음절 이외의 음절을 비표기 음절이라 정의한다. 표기 후보에서 완성형 한글에서 사용되지 않는 비표기 음절을 포함한 표기 후보를 검색하여 삭제하면 표기 후보 생성 과정에서 불필요하게 생성된 표기 후보의 수를 감소시킬 수 있다.



(그림 5) 비표기 음절 검색 구조도

(그림 5)는 비표기 음절 검색 전체 구성도를 나타낸 그림이다. 음절 복원 규칙 의해 생성된 표기 후보를 음절 분리한 후 검색 인덱스를 생성한다. 생성된 검색 인덱스로 완성형 테이블을 탐색하여 완성형 표기로만 이루어진 표기 후보에 대해서는 표기 후보로 유지하고, 완성형 표기 이외의 음절을 포함한 표기 후보를 삭제하여 표기 후보의 수를 감소한다.

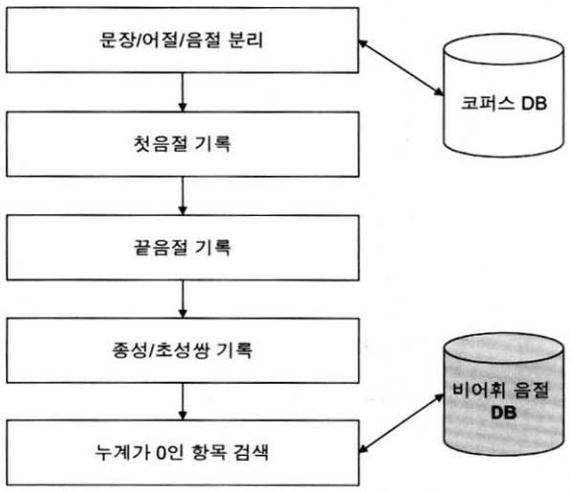
3.2 비어휘 음절 감소

한글에서 특정 음절은 제한된 어휘에만 사용되어 그 사용 빈도가 극히 적은 음절들이 있다. 국어 사전[18-21], 코퍼스 등에서 사용되는 어휘들을 검색하면 첫음절에는 쓰이지 않는 음절, 어절 내에서 사용되지 않는 음절과 종성 초성쌍 등이 발견된다. 본 논문에서는 이러한 첫음절과 끝음절에서 쓰이지 않는 음절을 비 어휘 음절이라 정의한다.

음절 복원 규칙을 통하여 생성된 표기 후보들 중에서 비어휘 음절과 종성 초성 쌍을 포함하는 표기 후보를 검색하여 삭제하면 형태소 분석 이전에 음절 복원 규칙 수를 감소시킬 수 있다.

3.2.1 비어휘 음절 검색 모듈

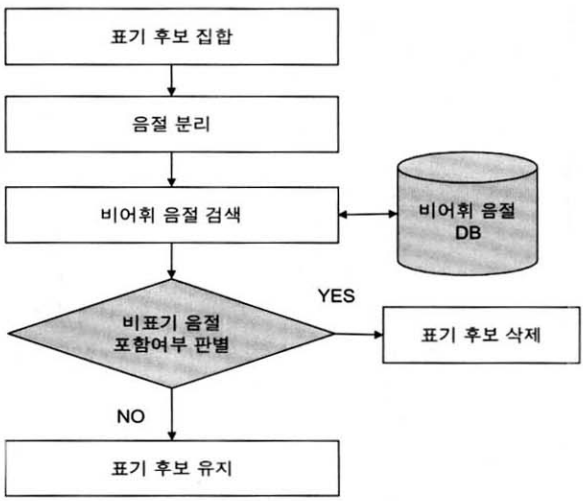
비어휘 음절을 검색하는 구조도를 (그림 5)에 나타내었다. 비어휘 음절 검색 모듈은 분야별로 수집된 대량의 코퍼스에서 첫음절과 끝음절에서 사용되지 않는 음절과 어절 내에서 나타나지 않는 어절쌍을 검색하여 비어휘 음절을 생성한다. 문장 분리는 코퍼스 내의 문장들을 마침표, 느낌표, 물음표 등의 문장 부호 단위로 분리한다. 어절 분리는 분리된 문장에서 스페이스, 텝 등의 화이트 스페이스와 따옴표, 쉼표 등을 구분자로 하여 어절로 분리하는 역할을 한다. 음절 분리는 어절을 각각의 음절로 분리한다. 분리된 음절에서 첫음절과 끝음절에 대하여 기록하고 음절쌍에 대하여 기록한다. 준비된 코퍼스에 대하여 모든 검색이 끝나면 누계가 0인 항목을 추출한다. 이때 추출된 항목이 비어휘 음절이 된다.



(그림 6) 비어휘 음절 검색 구조도

3.2.2 비어휘 음절 제거

음절 복원 규칙을 통하여 생성된 표기 후보에서 비어휘 음절 DB에 기록된 음절을 검색하여 제거하면 표기 후보의 수를 감소시킬 수 있다.



(그림 7) 비어휘 음절 제거 구조도

(그림 7)에 비어휘 음절 제거 모듈의 전체 구조를 나타내었다. 인식 음가열에서 음절 복원 규칙을 참조하여 표기 후보를 생성한 후에 비어휘 음절 제거 과정을 시작한다. 음절 분리 모듈에서 표기 후보의 어절중 하나를 선택하여 음절 분리를 수행하고 이어서 어절쌍을 생성한다. 생성된 어절과 어절쌍을 비어휘 음절 DB에서 검색하여 비어휘 음절이 검색되면 표기 후보를 삭제한다. 검색과정에서 첫음절, 끝음절, 종성 초성쌍 중 어느 하나라도 포함하고 있다면 그 표기 후보는 삭제된다.

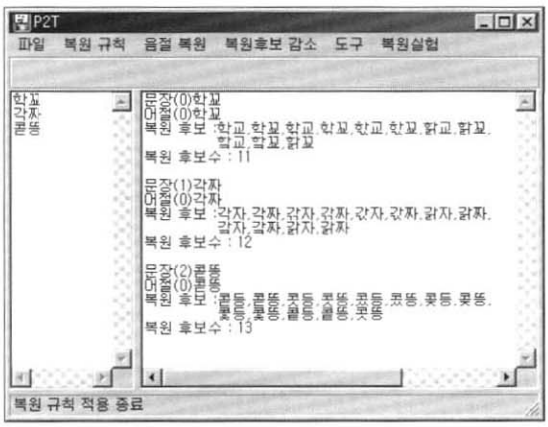
4. 실험 및 고찰

4.1 표기 후보 생성 실험

인식 문자열이 입력되면 표기 후보 생성모듈에서 음절 분리와 자모 분리를 거쳐 종성과 초성의 인덱스를 생성한 후, 표기 후보 생성 모듈에서 음절 복원 규칙을 참조하여 입력된 음가가 발음될 수 있는 모든 조합 가능한 표기를 생성해 낸다.

(그림 8)의 결과는 음가 '학꾜, 각짜, 콛뚱'을 직접 입력하여 각각의 음가로 발음되는 모든 표기 후보를 생성한 결과이고, (b)의 결과는 표준 발음법의 예제중 표준 발음법의 예외를 설명한 예제를 제외한 예제에 대한 실험 결과이다. 실험의 판정 방법은 표준 발음법에 표기된 단어와 발음 사전의 음가를 입력하였을 때 해당 음가의 표기가 포함된 표기 집합이 생성될 경우를 성공으로 판정하고, 올바른 표기가 생성되지 않은 경우를 실패한 경우로 판정한다.

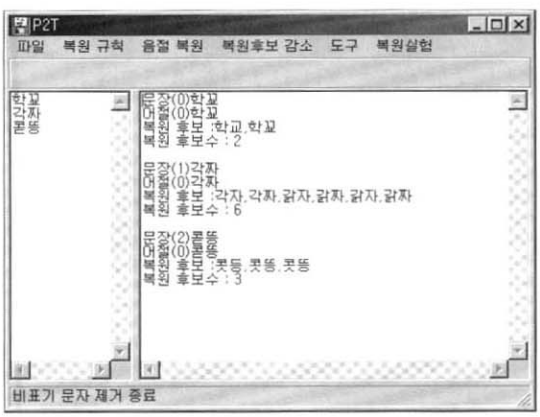
<표 1>은 표준 발음법 예제 195어절과 발음사전에서 임의로 발췌한 500어절에 대한 실험 결과를 요약한 표이다. 발음 사전에서 발췌된 500개의 어절의 발췌 기준은 음운 변동 현상이 나타나는 표제어 중 발췌된 어절과 동일한 형태의 음운 변동 현상이 나타나지 않은 예제만을 발췌하였다. 실험에 사용한 예제에서 발성이전의 표기를 포함한 표기 후보 집합이 생성됨을 알수 있다. 그러나 생성된 평균 후보수가 62개에 달하여 생성된 표기 후보의 수를 감소시키기 위한 방안에 대한 연구가 필요함이 실험 결과를 통하여 나타났다.



(a) 직접 입력에 의한 표기 후보 생성 결과

(b) 표준 발음법에 의한 표기 후보 생성 결과

(그림 8) 표기 후보 생성 결과



(a) (그림 8)(a)에 대한 실험 결과

(b) (그림 8)(b)에 대한 비표기 음절 감소 결과

(그림 9) 비표기 음절 감소 결과

〈표 1〉 표기 후보 생성 실험 결과 요약

실험 집합	어절 수	생성 성공 어절 수	평균 생성 후보수	생성 실패 어절 수
표준 발음법	195	195	65.05	0
발음 사전	500	500	59.23	0
평 균	347.5	347.5	62.14	0

4.2 비표기 음절 감소 실험

비표기 음절 감소 모듈은 표기 후보에서 비완성형 음절을 포함한 표기 후보를 검색하여 삭제하는 기능을 수행한다. 비표기 음절 감소 실험의 판정 기준은 감소 모듈을 수행한 후에 남아있는 표기 후보 집합에서 올바른 표기 후보가 남아있고, 비표기 음절을 포함한 표기 후보를 포함한 표기 후보가 없는 경우를 실험 성공으로 판정한다. (그림 9)는 (그림 8)의 결과에 비표기 음절 감소를 수행한 결과이다.

비표기 음절 감소시 모든 표기 후보가 삭제 대상이 되면, 원래 입력받은 음가를 유지하는 방법으로 비표기 어절 감소시 발생되는 문제점을 해결하였다. 이러한 방법으로 사람이나 사물의 이름과 고유명사, 외국어의 한글 음차에 대한 문제점을 해결할 수 있다.

<표 2>는 실험 집합에 대하여 비표기 음절 감소 실험 결과를 요약하여 나타낸 표이다. 실험 집합에 대하여 비표기 음절 감소 모듈을 수행한 후 판정 기준에 따라 성공과 실패를 분류한 결과 실험 집합 전체에 대하여 평균 34.8%의 표기 후보가 감소됨을 알 수 있었다.

〈표 2〉 비표기 음절 감소 실험 결과 요약

실험 집합	어절 수(개)	감소 성공 어절 수(개)	감소후 평균 후보수(개)	감소후 평균 감소율(%)
표준 발음법	195	195	40.86	31.45
발음 사전	500	500	36.45	38.21
평 균	347.5	347.5	38.65	34.83

4.3 비어휘 음절 감소 실험

비어휘 음절 감소 실험의 판정 기준은 한국 도서출판 중앙회의 '새 국어대사전'[18], 동아출판사의 '동아 새국어대사전'[19], 민중 출판사의 '국어대사전'[20], 금성출판사의 '뉴에이스 국어 사전'[21], KAIST 코퍼스를 대상으로 한 비어휘 음절 검색 결과에서 생성된 비어휘 DB에 기록된 단어가 감소 실험 결과에 존재하지 않은 것을 성공으로 간주한다.

비어휘 음절 감소 실험은 생성된 표기 후보중에서 비어휘 음절을 포함한 표기 후보를 삭제하여 전체 표기 후보수를 감소시키는 기능을 수행한다.

(그림 10)는 입력 음가열 '학교'에 대한 비어휘 음절 감소를 수행한 결과이다.

(a) (그림 8)(b)에 대한 비어휘 첫음절 제거 결과

(b) (그림 8)(b)에 대한 비어휘 중간 음절 제거 결과

(c) (그림 8)(b)에 대한 비어휘 종성 초성쌍 제거 결과

(그림 10) 비어휘 음절 감소 실험

〈표 3〉 비어휘 음절 감소 실험 결과 요약

실험 집합		어절 수(개)	감소 성공 어절 수(개)	감소후 평균 후보수(개)	감소후 평균 감소율(%)
표준 발음법	첫음절	195	195	40.86	31.45
	중간음절	195	195	22.60	44.31
	음절쌍	195	195	64.03	5.54
	평 균	**195**	**195**	**42.19**	**27.1**
발음 사전	첫음절	500	500	38.32	34.68
	중간음절	500	500	20.52	47.06
	음절쌍	500	500	62.02	7.11
	평 균	**500**	**500**	**40.24**	**29.61**
평 균		**347.5**	**347.5**	**41.24**	**28.36**

<표 3>는 실험 집합에 대하여 비어휘 감소 실험 결과를 요약하여 나타낸 표이다. 실험 집합에 대하여 비어휘 감소 모듈을 수행한 후 판정 기준에 따라 성공과 실패를 분류한 결과 실험 집합 전체에 대하여 성공함을 알수 있었다.

4.4 비어간 음절 감소 실험

(그림 11)은 비어간 음절 DB를 이용하여 표기 후보 중에서 비어간 음절이 포함된 표기 후보를 제거한 결과이다.

(그림 11) 비어간 음절 감소 결과

<표 4>는 표준 발음법 예제와 발음 사전 예제를 통하여 생성된 표기 후보 중에서 비어간 음절 제거 실험을 수행한 결과를 요약하여 나타낸 표이다.

〈표 4〉 비어간 음절감소 실험 결과 요약

실험 집합	어절 수(개)	감소 성공 어절 수(개)	감소후 평균 후보수(개)	감소후 평균 감소율(%)
표준 발음법	195	195	26.67	46.74
발음 사전	500	500	22.45	52.34
평 균	347.5	347.5	24.56	49.54

4.5 전체 감소 과정 적용 실험

(그림 12) 전체 감소 과정 적용 감소 결과

(그림 12)는 표준 발음법 예제에 대하여 음절 복원 후 비표기 음절 제거, 비어휘 첫음절 제거, 비어휘 중간 음절 제거, 비어휘 종성 초성쌍 제거, 비어간 음절제거를 차례로 수행한 결과이다.

<표 5>는 표준 발음법과 발음 사전에서 추출한 실험 집합으로 감소 과정의 전 과정을 적용한 후의 결과를 요약하여 나타낸 표이다.

〈표 5〉 감소 과정 전체 실험 결과 요약

실험 집합	어절 수(개)	감소 성공 어절 수(개)	감소후 평균 후보수(개)	감소후 평균 감소율(%)
표준 발음법	195	195	9.00	74.00
발음 사전	500	500	7.00	82.00
평 균	347.5	347.5	8.00	78.00

5. 결 론

본 논문에서 제안하는 음절 복원 규칙 생성 방법과 표기 후보 생성 및 감소 방안의 장점은 크게 네 가지이다. 첫째, 기존 규칙기반의 방법이 관찰에 의한 방법으로 음절 복원 규칙을 생성하여 규칙이 무결함을 증명할 수 없었던데 비하여 모델링을 통한 방법으로 변환 규칙을 생성하여 기존 관찰에 의한 방법의 문제점인 규칙의 무결성을 확인할 수 있다. 둘째, 자음과 모음에 대한 단 한번의 규칙 검색으로 기존 방법에서의 규칙 적용 알고리즘의 복잡성과 불필요한 반복을 문제 해결할 수 있는 장점을 갖는다. 셋째, 표기 후보 생성 면에서 복원 규칙에 따라 자모를 합성하여 표기 후보를 생성하므로 입력된 음가로 발성될 수 있는 모든 표기 후보가 생성됨을 보장할 수 있다. 넷째, 형태소 분석이전과 생성된 표기 후보의 감소를 통하여 형태소 분석과 구문 분석에 필요한 시간을 최소화할 수 있다.

본 논문에서는 표준 발음법에 제시된 표준 발음만을 고려하여 음절 복원 규칙을 생성하였으므로 수의적 음운 변동 현상은 고려하지 않았다. 추후 연구를 통하여 수의적 음운 변동 현상에 대한 처리까지 음절 복원 규칙에 포함하게 되면 인식 장치에서 인식된 음가에 대한 충실한 표기 복원이 이루어질 것이며, 본 논문에서 제안한 음절 복원 방안을 대화체 연속 음성 인식에 이용하면, 그동안 음성 인식 성능 향상에 걸림돌이었던 음운 변동 현상과 조음 효과에 대한 해결 방안이 될 것이다.

본 연구에서는 입력된 음가로 발음되는 모든 표기에 대한 표기 후보를 생성한 후에 과생성된 표기 후보를 감소시키는 방법을 사용하였다. 표기 후보를 생성한 후에 감소시키는 방법보다 생성 이전에 복원 규칙에서 과생성을 방지하도록 하면 더 빠른 음절 복원이 가능할 것이다.

참 고 문 헌

[1] 이행세, 음성 인식기법, 청문각, 1999.

[2] 이건상 외, 음성 인식, 한양대학교출판부, 2001.
[3] 박선호, 음성합성과 음성 인식시스템, 영진닷컴(주), 1990.
[4] 김병창, "형태소 그래프를 이용한 한국어 연속음성인식과 형태소 분석의 통합", 석사학위논문, 포항공과대학교, 1997.
[5] 심광섭, "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지(B), 제23권 제9호, pp.991-1000, 1996.
[6] 김계성, 이현수, 이상조, "연속 음절 문장에 대한 3 단계 한국어 띄어쓰기 시스템", 정보과학회논문지(B), 제25권 제12호, pp.1838-1844, 1998.
[7] 박미선, 김미진, 김계성, 최재혁, 이상조, "연속 음성 인식 후처리를 위한 음절 복원 rule-based 시스템과 형태소 분석기법의 적용", 전자공학회논문지, 제36권 c편 제3호, 1999.
[8] 서상현, "한글 음운 규칙에 기반한 음절 복원기 구현", 경북대학교, 석사학위 논문, 1997
[9] 이경님, 전재훈, 정민화, "한국어 연속 음성 인식을 위한 발음열 자동 생성", 한국음향학회지, 제20권 제2호, pp.35-43, 2001.
[10] 이원일, "신경망과 CYK-table을 이용한 음성 언어의 분석", 포항공대, 석사학위 논문, 1993.
[11] 임재걸, 이계영, 김경징, 김규식, "페트리넷을 이용한 표준 발음법 분석 시스템 구현", 한국정보처리학회 춘계학술발표논문집, pp.609-612, 1999.
[12] 임재걸, 이계영, 김경징 "표준 발음법 페트리넷을 이용한 음운 변환기 설계", 한국멀티미디어학회 춘계학술발표논문집, 제2권 제1호, pp.339-344, 1999.
[13] 이계영, 임재걸, 김경징, "표준 발음법의 일관성 검사와 우선순위 결정", 한국정보처리학회 추계학술발표논문집, 1999.
[14] Unicode Consortium, The Unicode Standard, Addison-Wesley Pub. Co, 1996.
[15] Graham, Tony, Unicode, John Wiley & Sons, 2000.
[16] 김성웅, "Tabular parsing 방법과 접속 정보를 이용한 한국어 형태소 분석기", 석사학위논문, 한국과학기술원, 1993.
[17] 박종만, "효율적인 한국어 형태소 분석기 및 철자 검사 교정기의 구현", 석사학위 논문, 서울대학교, 1990.
[18] 감수 이숭녕, 새 국어대사전, 한국 도서출판 중앙회, 1999.
[19] 동아출판사, 동아 새국어대사전, 동아출판사.
[20] 이희승, 국어대사전, 민중 출판사, 1988.
[21] 금성출판사 사서부, 뉴에이스 국어 사전, 금성출판사, 1989.

이 상 범

e-mail : sang107@dku.edu
1974년 연세대학교 전자공학과(공학사)
1978년 서울대학교 대학원 전자공학과 (공학석사)
1986년 연세대학교 대학원 전자공학과 (공학박사)
1984년 미국 IOWA대학교 컴퓨터공학과 객원교수
2000년 미국 SanJose대학 컴퓨터공학과 객원교수
1979년~1999년 단국대학교 전자·컴퓨터공학과 교수
1997년~1999년 단국대학교 교무·연구처장
1997년~현재 단국대학교 멀티미디어산업기술연구소장
2000년~현재 단국대학교 전기전자컴퓨터공학부 교수
관심분야 : 컴퓨터구조, 패턴인식, 디지털 신호처리

박 성 현

e-mai : mzzang@dankook.ac.kr
2000년 덕성여자대학교 불어불문학과 (문학사)
2002년 단국대학교 대학원 전자컴퓨터 공학과(공학석사)
2003년~현재 단국대학교 대학원 전자컴퓨터공학과(박사과정)
관심분야 : 자연어 처리, 음성 인식/합성, 디지털 신호처리