

최소 형태소 정보를 이용한 자동 발음열 생성 시스템

김 선 희*, 안 주 은**, 김 순 혁**

* 광운대학교 음성정보처리기술연구센터, ** 광운대학교 컴퓨터공학과

Automatic Pronunciation Generation System Using Minimum Morpheme Information

Sunhee Kim, Ju-Eun An, Soon-Hyob Kim

sunhkim@daisy.kw.ac.kr, dkswndms@hanmail.net, kimsh@daisy.kw.ac.kw

요약

본 논문은 최소한의 형태소 정보를 이용한 자동 발음열 생성 시스템을 제안한다. 일반적으로 발음열 생성 시스템은 입력된 문장에 대하여 형태소 단위로 분석한 다음, 각 형태소와 형태소의 결합 관계를 고려한 음운 규칙을 적용함으로써 상응하는 발음열을 생성한다. 지금까지의 연구는 이러한 발음열 생성시의 형태소 분석에 관하여 그 범위에 관한 연구 없이, 가능한 최대한의 분석을 상정하고 있다. 본 논문은 한국어 음운현상을 체계적인 텍스트 분석을 통하여 모든 형태론적 음운론적인 환경에서 가능한 모든 음운현상을 분류하여 발음열 생성시에 실제로 필요한 형태소 분석의 범위를 규명하는 것을 그 목적으로 한다. 음운 현상을 분석하기 위해 사용한 텍스트 자료로는 어휘가 충복되지 않으면서도 많은 종류의 어휘가 수록된 5만 여 어휘의 연세한국어사전과 2200 여 개의 어미와 조사를 수록한 어미조사사전을 이용하였다. 이와 같이 텍스트를 분석한 결과, 음운현상은 규칙적인 음운 현상과 불규칙적인 음운현상으로 나뉘는데, 이 가운데 형태소 정보가 필요한 형태음운규칙으로는 두 가지가 있으며, 이러한 형태음운규칙을 위한 형태소 분석의 범위로는 세세한 분류를 필요로 하지 않는 최소한의 정보로 가능함을 보인다. 이러한 체계적인 분석을 기반으로 제안하는 자동 발음열 생성 시스템은 형태음운규칙과 예외규칙, 그리고 일반음운 규칙으로 구성된다. 본 시스템에 대한 성능 실험은 PBS 1637 어절과 ETRI 텍스트 DB 19만 여 어절을 이용하여 99.9%의 성능 결과를 얻었다.

1. 서론

자동 발음열 생성이란 주어진 언어의 맞춤법 체계를 반영하고 있는 문자열을 음성 체계를 반영하는 발음열로 변환하는 것을 말한다. 이러한 자동 발음열 생성은 음성합성과 음성인식에 필수적인 요소 시스템으로, 철자표기에 대한 지식 및 음운론적인 지식이 그 토대가 된다.

기존의 표기음성변환을 위한 자료로는 “표준어 규정” 제2부의 표준 발음법과 발음 사전들([1], [2])이

이용되어 왔는데, 먼저, 표준발음법은 우리말에서 모든 가능한 음운현상을 포함한 것이 아닌 부분적인 기술에 그치고, 발음사전은 그 발음현상에 대한 전체적인 체계보다는 낱말의 개별적인 발음 표기를 제시한 것으로 실제 표기음성변환을 위한 직접적인 자료로서는 불충분하다.

본 연구는 우리말 맞춤법과 표준 발음법을 고려하고, 텍스트 코퍼스 분석을 통하여 우리말의 철자 체계에서 발음 체계로의 변환을 위해 필요한 형태-음운론적 지식

을 체계적인 텍스트 분석을 통하여 정리한다[3]. 분석 대상이 되는 텍스트 자료로는 우리말에서 사용되는 모든 가능한 음소연결과 그 음운현상을 분석해 내기 위하여 5만여 어휘를 수록한 연세한국어사전[4]과 2200여 개의 어미와 조사를 수록한 어미조사사전[5]을 이용하였다.

이와 같이 텍스트를 분석한 결과, 음운현상은 규칙적인 음운 현상과 불규칙적인 음운현상으로 나뉘는데, 가운데 형태소 정보가 필요한 형태음운규칙으로는 두 가지가 있으며, 이러한 형태음운규칙을 위한 형태소 분석의 범위로는 기존의 연구[6]와는 달리 세세한 분류를 필요로 하지 않는 최소한의 정보로 가능함을 보인다.

논문의 순서는, 먼저, 텍스트 분석을 통한 음운현상을 체계적으로 정리한 결과를 제시하고, 이를 기반으로 한 한국어 자동 발음열 생성 시스템을 제안한 다음[7], 제안한 자동 발음열 생성 시스템의 성능 실험 결과를 보기로 한다.

2. 텍스트 분석 기반 한국어 음운현상

음운현상은 음운환경에 따라 자음간의 음운현상, 자음과 모음간의 음운현상, 모음간의 음운현상, 모음과 모음간의 음운현상으로 나누어 볼 수 있다. 이렇게 분류되는 음운현상은 다시 어절 내부와 어절 경계로 나누어 볼 수 있는데, 본 논문은 일단 어절 내부의 음운현상으로 한정하고, 또한 그 가운데 수의적인 현상은 제외한다.

어절 내부에서 관찰되는 음운현상은 다시 크게 단어내부, 체언과 조사의 경계, 동사 어간과 어미의 경계 등의 세 가지로 분류하여 살펴 볼 수 있다. 여기에서 단어 내부란 일반 사전에서 표제어로 등재된 모든 어휘를 의미하여, 분석대상인 연세한국어사전에 수록된 5만여 개로 이루어 진다. 여기에서 사전의 표제어는 복합어와 파생어를 포함하여 많은 합성어가 포함된다.

다음 <표 1>은 텍스트 자료를 분석하여 음운현상을 정리한 것이다. 음운현상은 크게 규칙적인 음운현상과 불규칙적인 음운현상으로 나뉘는데, 규칙적인 음운현상은 다시 일정한 음소 문맥이 주어졌을 때 항상 관찰되는 일반음운현상과, 일정한 음소 문맥이 일정한 형태소

경계에 주어진 경우에만 관찰되는 형태 음운현상으로 나누어 진다.

규 칙 적	일반	종성중화(1)
		자음군 단음화(2)
		유음의 비음화(3)
		음운론적 경음화(4)
		격음화(5)
		장애음의 비음화(6)
		유음화(7)
		융합(8)
		이중 비음화(9)
		'ㅎ' 탈락(10)
		'ㅎ' 비음화(14)
		연음(21)
		구개음화(22)
		이중모음의 단모음화(23)
	형태	형태론적 경음화(13)
불규칙적		어휘적 경음화(11)
		유음화 예외(12)
		'ㄴ' 침가(24)
		중화(단순화)+연음(25)

<표 1> 텍스트 분석 기반 음운현상의 분류

여기에서 일정한 음소 문맥이 일정한 형태소 경계에 주어진 경우에만 관찰되는 음운현상을 일반 음운현상과 대비하여 형태음운현상이라 지칭하였는데, 이러한 음운현상은 동사어간과 어미사이에서만 관찰된다. 형태음운현상이 관찰되는 환경을 정리하면 다음 <표 2>과 같다.

	ㄷ	ㅅ	ㅈ	ㄱ	ㄴ
ㅁ	13	13	13	13	
ㄴ	13	13	13	13	
ㄉ	2	13	13	13	
ㄪ	2	13	13	13	
ㄮ	2	13	13	13	7

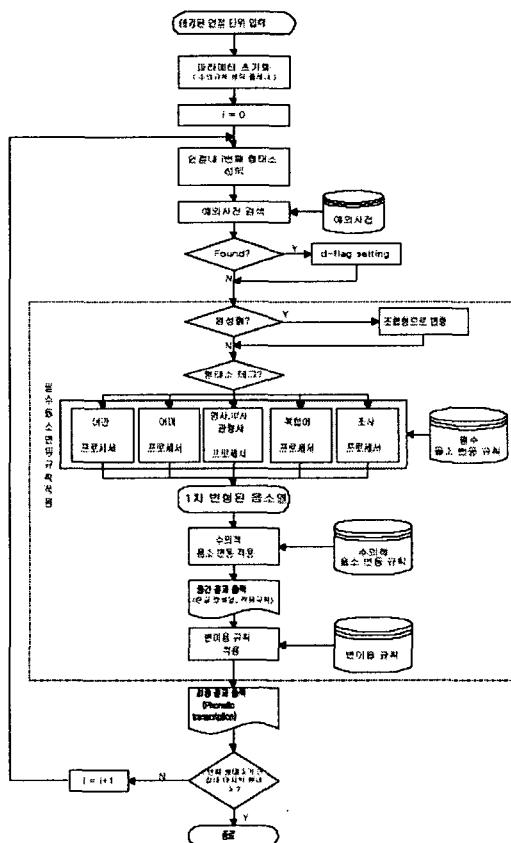
<표 2> 형태음운현상

(2) 자음군 단순화 (7) 유음화 (13) 형태론적 경음화

형태음운현상으로는 형태론적 경음화가 있는데, 이는 동사어간의 종성이 유성자음이나 유성자음을 포함한 자음군인 경우에 어간의 초성이 평음인 경우, 이 평음이 경음으로 실현되는 경우이다. 형태론적 경음화가 동사어간과 어미 사이에서만 관찰된다는 사실은, 자동 발음열 생성 시스템의 개발에 있어서 일반적으로 선행되는 형태소 분석의 범위를 한정해 주는 단서로서, 자동 발

음열 생성 시스템에 필요한 형태소 분석은 등사 어간과 어미를 분석해 주는 최소한의 분석으로 충분하다고 할 수 있다.

기존의 연구[6]는 다음 <그림 1>에서 보는 바와 같이, 체언과 조사, 동사 어간과 어미 뿐만 아니라, 복합어의 경우도 모두 형태소 분석을 상정하여 분석하는 것을 원칙으로 하고 있으나, 본 논문에서는 이와 같은 복합어와 파생어에 대한 세부적인 분류는 필요하지 않고, 이러한 부류의 단어들을 불규칙한 음운현상인 어휘적 경음화를 보이는 경우로 분류하여, 자동 발음열 생성 시스템에서 예외발음사전을 구성하도록 한다.



<그림1> 기존 혈태소 분석에 기반한 박을열 생선 시스템

예를 들어, /감기/ 라는 단어는 두 가지의 다른 형태 소 분석 결과를 나타낼 수 있는데, 첫 번째로, 동사구로 판단되며 어간인 /감/ 과 어미인 /기/로 되어, /감끼/로 발음된다. 반면에, 명사일 경우 /감기/로 발음되므로, 이와 같은 경우는 형태소 분석이 필수적이다.

선행되어야 한다. 그러나 같은 명사인 경우에 형태소 정보가 같지만 다르게 발음될 수가 있는데, 예를 들면, ‘임금[임금]’과 ‘갈립길[갈립길]’로 형태소 분석의 결과가 실제 발음을 예측하는 데 아무런 도움이 되지 않는다.

이렇게 불규칙한 경음화를 보이는 예로는 고유어의 복합어 뿐만 아니라, 많은 한자어의 경우를 포함한다. 한국어의 어휘에 있어서 많은 부분을 차지하고 있는 한자어의 경우는 각각의 글자가 개별 형태소와 같은 행태를 보여 복합어에서와 같은 경음화 현상이 관찰되나, 이러한 경음화 현상은 고유어 복합어의 문제 보다 규칙화 하는 것은 더욱 어려운 일이다. 본 논문에서는 이러한 한자어는 복합어와 마찬가지로 그 형태소 분석을 배제하고 예외별음을 보이는 단어로 분류한다.

이러한 불규칙적인 음운현상은 자음과 자음간, 자음과 모음간, 그리고, 모음과 자음간에 관찰되는데, 그러한 각각의 경우에 해당하는 음운현상은 다음 <표 3>와 같다.

자음 + 자음	어휘적 경음화 유음화 예외
자음 + 모음	'ㄴ' 침가 중화(단도음화) + 연음
모음 + 자음	어휘적 경음화

<표 3> 불규칙적인 음운현상

어휘적 경음화란 비음 뒤에 평음이 오는 경우와 모음 다음에 평음이 오는 경우에, 이 평음이 경음화되는 현상을 의미한다. 유음화의 예외란 일반적으로 유음화가 관찰되는 환경(ㄴ+ㄹ)에서 일부 단어의 경우 예외적으로 유음(ㄹ)이 비음(ㄴ)으로 실현되는 것을 의미한다.

다음 표는 예외발음이 관찰되는 환경을 정리하여 음영으로 나타내고, 각각의 경우의 예를 들었다.

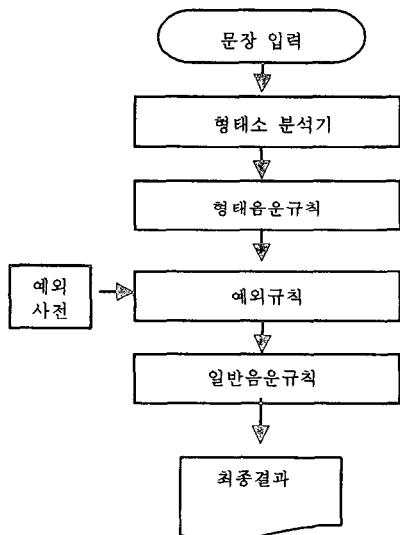
	ㅂ	ㄷ	ㅅ	ㅈ	ㄱ	ㄹ	ㅇ
ㅁ	봄비	보름달	술소리	홈집	꿈길		
ㄴ	눈병	눈독	눈살	관점	눈가	법신론	
ㅇ	嘭별	장독	망속	동적	땅글		
ㄹ	들보	갈대	결산	결재	들것		
ㅂ	육아법	대가백	여마살	제점	내파		
c							웃웃 앞일

<표4> 예외발음환경과 그 예

3. 자동 발음열 생성 시스템

본 논문에서 제안하는 자동 발음열 시스템은, 위에서 살펴본 규칙적인 음운현상인 일반음운현상과 형태음운현상을 각각 일반음운규칙과 형태음운규칙으로 규칙화하고, [5]에서 불규칙적인 음운현상을 보이는 어휘들을 추출하여 예외사전과 예외규칙을 만들어, <그림 2>와 같이 형태음운규칙, 예외규칙(예외사전 검색), 일반음운규칙의 순서로 적용한다.

전처리 된 입력 문장은 형태소 분석기를 통해 형태소로 분석되는데, 이 때 어간과 어미 정보를 가지는 동사구에만 형태음운규칙인 형태론적 경음화가 적용된다.



<그림 2> 발음열 자동 생성 알고리즘

형태음운 규칙이 적용된 어절을 제외한 나머지 어휘들 중에, 예외 규칙이 적용될 어휘를 선별하는데, 이때, 예외사전을 검색하여 예외사전에 있는 어휘들에게만 예외규칙을 적용한다. (여기에서 예외 사전은 텍스트 분석을 통하여 2,931 개의 예외 단어로 구성되었다.) 마지막으로, 규칙적 음운 현상이 적용될 수 있는 어휘들에 일반음운규칙을 적용한다.

4. 실험 및 결과분석

실험을 위한 텍스트로는 시스템 공학 연구소(SERI)의 PBS 1637 어절과 한국전자통신연구소(ETRI)의 신문

사설 텍스트 19,2385 어절을 사용하였다. 실험 결과, 99.9 % 의 정확도를 나타냈으며, 전체 시스템에서 나타난 0.1%의 오류는 불규칙 음운현상 중에 하나인, 어휘적 경음화 현상에 기인한 것이다. 전체 시스템에서 형태음운규칙이 차지한 비중은 0.1%에 해당하였다.

5. 결론

본 논문은 텍스트를 분석하여 음운 현상을 체계적으로 정리한 결과, 동사의 어간과 어미 정보만을 나타내는 최소 형태소 정보만을 이용하여 보다 체계적인 자동 발음열 생성 시스템을 구현할 수 있었다. 실험 결과, 많은 양의 텍스트에서 형태음운규칙이 적용되는 비중이 매우 적었고, 이를 처리하기 위하여서는 하나의 형태음운규칙으로 충분하였다. 이와 같이, 자동 발음열 생성 시스템을 개발함에 있어서는 최소한의 형태소 정보만을 이용함으로서 충분히 좋은 성능을 얻을 수 있음을 보였다.

참고문헌

1. 김석득, 이현복, 유재원, 표준 한국어 발음 대사전, 어문각, 1993.
2. 이현복, 한국어 표준발음사전, 서울대학교 출판부, 2002.
3. 김선희, “한국어 자소열-음소열 변환을 위한 음운 현상 연구”, 한국음향학회 하계학술발표대회 논문집, 2003.
4. 연세대학교 언어정보연구원, 연세한국어사전, 두산동아, 1998
5. 이희자, 이종희, 어미조사사전, 한국문화사, 2001
6. 전재훈, “형태음운학적 분석에 기반한 한국어 발음 열 자동 생성”, 서강대학교 석사학위논문, 1997
7. 김선희, 안주은, 김순협, “텍스트 분석 기반 한국어 자동 발음열 생성 시스템”, 음성신호처리학회, 2003