표준어 규정에 따른 한국어 음소열 자동생성기

이도관*, 강미영, 윤근수†, 이교운†, 권혁철
부산대학교 전자계산학과
{dklee*, kmyoung, hckwon}@pusan.ac.kr, {ksyun, kwlee}†@mail.ulsan.ac.kr

Automatic Phoneme Generator based on Standard Korean Pronunciation

Do-Gwan Lee*, Mi-Young Kang, Geun-Soo Yoon†, Gyo-Un Lee†, Hyuk-Chul Kwon
Dept. of Computer Science, Pusan National University, Dept. of Computer Information, Ulsan College†

요 약

우리말에서 띄어쓰기와 버금갈 정도로 어려운 것이 우리말의 발음이다. 이에 실생활에서 혼란스럽게 사용되는 발음법과 그로 인해 올바른 발음의 선택에 대한 어려움을 덜어낼 수 있도록 표준어 규정의 표준 발음법에 따른 한국어 음소열 자동 생성기를 구현하여 교육용으로 쓸 수 있도록 하는 것이 이 논문의 목적이다.

1. 서 론

모든 언어는 끊임없이 변한다. 이런 변화로 한국어 또한 어법 상의 혼동이 생기는 경우가 많이 생기는데 이에 대한 규범을 제시하기 위해 표준어 규정이 1988.1.9 문교부고시 제88-2호를 따라 제정되었다. 그러나 실제 언어사용에서 어문규범이 경시되거나 잘못 쓰이는 경우가 흔하다. 특히, 띄어쓰기의 경우는 모국어 화자인 한국인에게도 어렵기 때문에 잘못 쓰인 예가 많은데 맞춤법 검사기가 상당한 정도로 이 부분을 보완한다[1]. 또한, 발음 혼동이 일어나는 경우가 많다. 발음 기호 없이도 읽을 수 있고, 소리와 글자의 대응 관계를 알면 쉽게 적을 수 있는 표음 문자임을 자랑하는 한국어의 모국어 화자들이 주어진 쓰인 텍스트에 대한 발음에 혼동을 일으켜 오류를 범하는 것은 큰 문제이다.

이와 같이 우리말의 발음에서 많은 혼동이 일어나는 데는 우리말의 역사적 특성상 한자어가 많으며, 이로 인해 한자어와 관련된 동철이음어(同綴異音語)가 많다는 사실이 크게 작용한다. 음운 변화는 연결 제약 조건, 어절 내 구성 형태소의 종류와 연결 형태에 의존하는데[2] 동철이음어의 경우는 문법적 의미와 어휘적 의미의 차이에 따라 다른 발음을 가질 수 있어 의미를 고려하지 않으면 구성 형태소의 종류와 연결 형태를 정확하게 밝혀내는 것이 어렵다[3][4]. 이와 같은 발음 혼동에 대한 규범을 제시하기 위해 종래에는 없었던 것이지만 '표준 발음법' 규정이 새로이 제정되었다. 그러나 우리 말의 발음 중에서 예외적으로 발음되는 단어가 많다. 전체 표준발음법 규정 30개 항 중, 11개 항에서 총 17개의 예외 규정이 있으며, 각 항에 대한 추가 규정이라고 할 수 있는 [붙임]규정은 12개 항에 총 15개가 있다. '표준 발음법'에 모든 혼동을 해결해줄 수 있는 충분한 규범을 제시하고 있지는 않다. 따라서 문법적 의미와 어휘적 의미의 차이에 따른 발음 차이를 보일 수 있는 발음 표현 생성기의 개발이 필요하다.

2장에서 음소열 생성기 관련한 중요한 문제점을 중심으로 기존 연구에 대해 알아보고 3장에서 이 논문이 제안하는 발음 표현 생성기 구조를 설명하고, 앞서 말한 한국어 발음 문제를 해결하기 위해 이 논문이 개발한 부분문장 분석을 이용한 의미 태깅 기법을 알아본다. 4장에서는 실험 결과를 정리하며, 끝으로 5장에서 이 논문의 전체 결과 및 향후과제를 정리한다.

2. 기존 연구

정확한 한국어 발음 표현 생성기를 구현하기 위한 이전 연구들을 살펴보면, 동철이음어와 관련한 문제를 처리하기 위해 연결 제약 조건과 어절 내 구성 형태소의 종류와 연결 형태 2가지를 고려하여 연음법칙, 구개음화, 경음화 등의 규칙으로 분류하고, 다시 그 규칙들을 적용양상에 따라서 세부규칙으로 나누기도 하였고[2], 규칙을 단순화해서 표준 발음 규칙이 적용되는 경계점의 앞뒤 형태소가 실질 형태소인지 아닌지만 구분하기도 하였다[5].

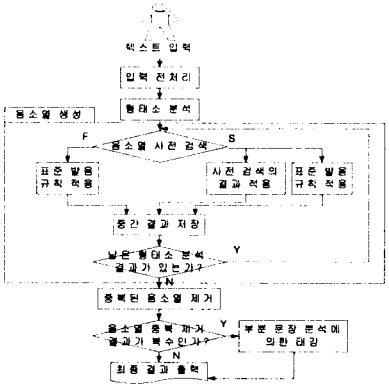
예외적 발음 단어 문제에 대해서는 예외사전을 구축하여 해결하고 있다[2]. 그리고 음성 패턴 사전(Phonetic Pattern Dictionary)과 형태 음소 연결 표(Morphophonemic Connectivity Table), 그리고 CCV(Consonant Consonant Vowel) LTS(Letter to Sound) 규칙들을 이용하여 위 2가지 문제에 해결을 제안하기도 하였다[6]. 하지만, 이전 연구들은 의미에 기반한 태깅을 하지 않았기 때문에 동철이음어를 처리하는 데 있어서는 한계를 보이고 있다.

기존 연구에서는 음성 인식률의 향상에 목적을 두었기 때문에 우리말의 음소 목록에는 없다 하더라도 언어사용 수준에서 만날 수 있는 발음열도 생성하였다[2]. 그러나, 이 논문의 목적은 실제 언어사용에서 혼동을 일으키는 발음과 그로 인해 올바른 발음의 선택에 대한 어려움을 덜어내는 데 있기 때문에, 표준발음법을 토대로 음운론적 차원에서만 발음표기를 생성하여 보여준다. 국제음성기호(IPA)를 이용한 변이음을 통합하는 발음표현 시스템도 현재 구축중에 있다.

3. 음소열 자동생성기

임의의 입력 단어에 대한 정확한 발음표현을 생성하기 위해서는 언어학적 여러 단계(형태음운론, 통사-의미론)에 대한 이해를 바탕으로 형태소 내적 발음 규칙, 형태소 간의 발음 규칙, 구 단위의 발음 규칙을 반영한 발음표현 시스템이 필요하다. 이러한 요구에 부응하기 위하여 개발한 음소열 자동생성기의 구조를 각 모듈별로 아래에 설명한다.

3.1 음소열 자동생성기의 구조



[그림 1] 음소열 자동생성기 flowchart

3.1.1. 입력 전처리

입력은 단어나 문장이 될 수 있으나 한글에 대해서만 처리한다. 즉, 기호·단위, 숫자처리, 외국어 등은 이 논문에서 제외되었는데 기호·단위, 숫자처리는 현재 본 연구실에서 개발하고 있으며 외국어에 대한 음소열 생성은 추후의 연구에서 보강될 것이다. 단, 한자인 경우는 한글로 변환해서 처리한다. 그래서 이 연구에서 정확도를 측정할 때에는 한글이 아닌 다른 문자를 포함하는 어절은 제외한다.

3.1.2. 형태소 분석

음소의 변화는 형태소 분석 결과에 의존하므로 형태소 분석기의 성능에 따라 정확도가 크게 좌우될 수 있다. 이 시스템에서 사용하는 형태소 분석기는 본 연구실에서 개발한 것을 사용하였고, 신문 기사에 대해 미등록어와 입력오류 등을 포함한 형태소 분석기의 정확도는 95.79%이다[7].

3.1.3. 음소열 생성

음소열 생성을 위해서 먼저 음소열 사전을 검색하는데, 음소열 사전 구축에 대해서는 3.2장에서 자세히 설명하겠다. 입력된 단어에 대해 음소열 사전 검색이 성공할 경우는 음소열 사전의 음소열 결과뿐만 아니라 형태소 분석 결과를 참조해서 표준 발음규칙도 적용을 한다. 음소열 사전 검색에 실패한 단어는 표준 발음 규칙만 적용한다. 형태소 분석에 실패한 단어 즉, 미등록어인 경우도 마찬가지로 수행되지만 형태소 분석 정보가 없기 때문에 일반적인 음운규칙만 적용된다. 또한, 앞 어절의 마지막 종성에 대한 정보와 앞 뒤 단어들의 품사 정보도 가지고 처리하기 때문에 한 통사적 구(phrase) 안에 위치하는 어절 간에 영향을 주는 경우도 처리한다.

(1) '멀어져 갈 것이다' -> [머러저 갈 꺼시다] [제 27항]

한 어절에서 생성될 수 있는 음소열의 총 개수(N_T)는 다음과 같다.

$$N_T = N_i \ (if\ N_{ma} = 0),\ N_T = \sum_{i=1}^{N_{ma}} N_i \ (if\ N_{ma} > 0)$$

(N_{ma}: 형태소 분석 결과 개수, N_i: 한 형태소 분석에서 생성된 음소열 수 - 사전 검색 실패시 1, 사전 검색 성공시 2)

3.1.4. 중복된 음소열 제거

형태소 분석을 마친 후의 음소열은 총 N_T개 만들어지지만, 형태소 분석 정보가 다를지라도 음소열이 중복되는 경우가 있을 수 있다. 이 단계에서는 그 중복되는 음소열들을 제거한다. 중복된 음소열을 제거한 후 음소열의 총 개수(N_r)은 다음과 같다.

$$N_r = n \ (1 \le n \le N_{ma})$$

3.1.5. 부분 문장 분석에 의한 태깅

중복된 음소열 제거 단계를 거친 후 $N_r > 1$이면, 이러한 동철이음어를 태깅하기 위한 의미에 기반한 문장 분석이 필요하다. 이 논문은 시스템의 부하를 줄이기 위하여 기 개발된 부분 문장 분석(partial parsing) 기법[8]을 이용한다. 예를 들어 '신고'와 같은 형태는 다음과 같은 형태소 분석 결과를 보인다.

[표 1] 형태소 분석과 음소열 생성

입력	형태소 분석	음소열
신고	한자어 명사(2)	신고
	규칙 형용사 + 어미(3)	신고
	규칙 동사 + 어미(4)	신꼬

중복된 음소열을 제거하면 [신고]로 발음 나는 경우와 [신꼬]로 나는 경우로 축약된다. 문제의 입력단어가 다음과 같은 문장 속에서 나타나는 경우는 부분 문장 분석을 통해서 정확한 음소열을 하나 태깅해야 한다.

(2) 출생 **신고**를 하러가자. (3) 과일이 **신고**? <--[신고]
(4) 신발을 **신고** 기다렸다. <---[신꼬]

우선, 중복된 음소열 제거 단계를 거친 후 $N_r > 1$의 음소열을 보여주는 '신고'는 부분 문장 분석 모듈로 넘어가게 되는데 이를 출발점으로 품사 및 의미분류 정보에 기반하여 '신고'와 함께 쓰일 수 있는 어절이 있는지 검색한다. 예를 들어 문장 (4)에서 '신고'를 출발점으로 하여 부분 문장 분석을 하면 '신발을'이라는 어절을 찾게 된다. 이 어절에서 '신발'은 '입고 신을 수 있는 것'이라는 의미분류 정보를 가지고 있는 어절이다. 따라서 '신고'라는 입력 어절의 남아있는 가능한 발음 중에서 동사 (4)에 해당하는 발음 표현을 태깅하게 된다. 이러한 기법은 사실상 문법적 의미 차이는 없고 어휘적 의미 차이만을 보이는 '안벽' 과 같은 예에 대한 태깅에서 더 유용하다.

[표 2] 형태소 분석과 음소열 생성

입력	형태소 분석 / 의미 분석	음소열
안벽	명사(5) / 깎아지른 듯이 험한 물가.	안벽
	명사(6) / 건물의 안쪽에 있는 벽.	안뻑

(5) 바닷가의 안벽을 봐라.

'안벽'이란 단어를 포함하는 문장이 입력이 되면 여러 가지 가능한 발음표현중 해당 발음을 태깅하기 위하여 문제의 단어를 중심으로 부분 문장 분석을 하여 '바다'나 '강'과 '물가' 같은 양립관계에 있을 수 있는 의미분류정보를 가진 어절이 있는지 검사한다. 문장 (5)에서는 이런 의미정보를 포함하는 어절이 있으므로 명사 (5)를 태깅하게 된다.

3.1.6. 최종 결과 출력

사용자들에게 올바른 발음을 선택하도록 교육하는 것이 이 시스템의 목적이므로 쉽게 이용하고, 접근할 수 있어야 하며, 결과도 이해하기 쉬워야 한다. 그래서 최종 결과 출력시 각 형태소 분석 별로 해당 음소열도 함께 출력하고 각 음소열 생성시 적용된 표준 발음 규칙들을 함께 보여줌으로써 교육적인 효과를 높였다.

3.2 음소열 사전

예외적으로 발음되는 단어들을 직접 찾아서 예외 사전을 구축한다는 것은 많은 시간과 작업량을 요구한다. 이 논문에서는 국어사전이 일반적인 음운규칙을 따르지 않는 단어들에 대해서는 그 음소열을 따로 표기한다는 사실에 착안해서, 국어사전에서 예외로 제시하는 단어들을 재분석하여 음소열 사전을 구성하였다. 그러나 음소열 생성규칙에 따라 자동 생성될 수 있는 많은 단어들의 발음이 국어사전에는 표시되어 있다. 예를 들어 '각주'와 같은 단어는 한국어의 일반 음운규칙인 경음화를 반영하는 표준 발음법 제 23항에 의거 [각쭈]로 발음될 수밖에 없지만 국어 사전에는 이런 단어도 발음 표시되어 있는데 재분석 결과를 바탕으로 이와 같은 단어의 발음은 음소열 사전에서 제외한다. 또한, 각 단어의 의미 및 형태론적 정보분류 정보(현 시스템은 약 160개 정도의 단어 분류 정보를 가지고 있다)를 바탕으로 규칙처리가 가능한 것도 자동 생성할 수 있기 때문에 제외한다. 예를 들어 '보문로'와 같은 형태는 표준발음법 20항 [붙임]에 따라 [보문노]로 발음된다. 이에 반해서 '천리'와 같은 형태는 표준발음법 제 20항의 적용을 받아 [철리]로 발음된다. 즉, '...ㄴㄹ...' 자음 연속이 있는 단어들이 [..ㄹㄹ..]로 발음나는 것과 [..ㄴㄴ..] 발음나는 것으로 분류되는데 이들은 명사+접미사로 형태소분석되는 단어들과 그렇지 않은 단어들로 분류될 수 있어 규칙으로 자동 생성이 가능하므로 음소열 사전에서 제외한다.

4. 실험 및 결과

실험에 쓰인 자료는 소설 1편과 신문 자료이며, 어절은 348,233개이다. 전체 자료에서 N_r > 1인 어절에 대해 조사를 해보았는데 결과는 [표 3]과 같다.

[표 3] 중복 음소열 제거 실험결과

분석 결과 / 분석 대상	개수	전체 어절에 대한 비율(%)
N_{ma} > 2인 어절	341,268	97.99
N_r > 1인 어절	6,859	1.97

그리고 음소열 생성의 정확도의 결과는 [표 4]와 같다.

[표 4] 음소열 자동생성기의 정확도

분석 결과 / 분석 대상	오류 개수	정확률(%)
부분 문장 분석	195	97.16
음소열 생성	454	99.87
전체	649	99.81

위 오류들의 유형을 조사한 결과를 [표 5]에 정리하였다.

[표 5] 오류 유형 분석

오류 유형	오류 개수	전체 오류에 대한 비율(%)
미등록어, 합성어	303	46.69
어절 간	217	33.44
복합명사	105	16.18
형태소 분석 오류	13	2.00
발음이 예외적인 단어	11	1.69

5. 결론 및 향후 과제

이 논문은 정확한 음소열 자동생성기를 구현하기 위하여 음소열 사전 구축과 중복 음소열 제거 등의 방법을 제안하였으며, 동철이의어를 문법적 의미나 어휘적 의미에 기반해서 처리할 수 있도록 문장 분석에 의한 태깅 방법도 제안하였다. 이 시스템의 성능 측정을 한 결과 이전 연구[5]보다 1.02%향상된 99.81%라는 높은 정확률을 나타내었다. 이 논문에서는 한글인 어절만 처리하였기 때문에 기호·단위 및 숫자처리를 완성하고 외국어 및 고유명사 변환에 대한 처리가 보강되어야 하겠다. 그리고 규칙 처리 루틴과 음소열 사전을 다듬어서 실험 중 발견된 예외 발음 단어에 대한 처리도 가능하게 해야 할 것이다.

<Acknowledegement>

이 논문은 국가지정연구실사업(과제명: 언어 중심의 지능적 정보처리를 위한 단계적 우리말 분석기술의 개발 (M10203000028-02J0000-01510))의 지원을 받아 이루어진 것임.

참고 논문

[1] H.C.Kwon, Y.S.Chae, D.I.Park, "A Practical Speller Using Multiple Dictionaries and a Corpus", Proceeding of IASTED, pp. 204-207, 1997.

[2] 이경님, 전재훈, 정민화, "한국어 연속음성 인식을 위한 발음열 자동 생성", 한국음향학회지, 제 20권, 제 2호, pp. 35-43, 2001.

[3] 이상호, 오영환, 서정연, "한국어 문서 음성 변환 시스템을 위한 문서 분석기", 한국음향학회지, 제 15권, 제 3호, 1996.

[4] 홍성훈, 전범기, 홍준모, 이주헌, 임재열, 안수길, 임홍순, "한국어 문장-음성 변환기의 언어처리에 관한 연구", 제 10회 음성통신 및 신호처리 워크샵 논문집, pp. 99-103, 1993.

[5] Do-Gwan Lee, Hyuk-Chul Kwon, "Automatic String Generator based on Standard Korean Pronunciation", APIS II, Jakarta, Indonesia, pp. 47-51, 2002.

[6] Byeongchang Kim, Gary Geunbae Lee, Jong-Hyeok Lee, "Morpheme-Based Grapheme to Phoneme Conversion Using Phonetic Patterns and Morphophonemic Connectivity Information", ACM TALIP, Vol.1, No.1, pp. 65-82, 2002.

[7] 김민정, "규칙과 말뭉치를 이용한 한국어 형태소 분석과 중의성 제거", 부산대학교 전자계산학과 박사학위 논문, 1996.

[8] M.Y. Kang, S.H. Park, A.S. Yoon, and H.C. Kwon, " Potential Governing Relationship and a Korean Grammar Checker Using Partial Parsing", IEA/AIE, Cairns, Australia, LNAI 2358, 2002.