1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

음성합성을 위한 불규칙 발음 사전 구축 방법

유재원
한국외국어대학교 언어학과

A design of a Korean Irregular Pronunciation Dictionary for Speech Synthesis

Yu Jaewon
Department of Linguistics
Hankook University of Foreign Studies
jwyu@maincc.hufs.ac.kr

요약

이 논문은 우선 한국어 음성 합성에 있어 불규칙 발음 사전이 왜 필요한가를 살피고, 이어서 '표준 한국어 발음 대사전(1993 한국방송공사 편찬)'을 바탕으로 어떻게 불규칙 발음 사전을 구축하는 것이 좋은가에 대하여 논의한다.

1. 불규칙 발음 사전의 필요성

문자로 된 문서를 말소리로 바꾸기 위해서는 우선 맞춤법에 맞춰 적은 낱말들에 변동 규칙을 적용하여 발음 나는 대로 적은 뒤, 이 음소들에 다시 변이음 규칙을 적용하여 환경에 따른 발음의 미세한 차이를 밝히고, 이어서 각 미세한 발음에 대응하는 음성 데이터를 찾아 기계로 소리를 합성한다.

이 과정에서 가장 어려운 문제는 변동 규칙으로는 이끌어 낼 수 없는 불규칙한 발음을 갖는 낱말들의 처리이다. 우리말에서 불규칙한 발음의 원인이 되는 현상들은 아래와 같다.

1) 모음의 장단: 한국어의 표준말은 낱말 첫 모음의 길이에 따라 뜻이 달라진다. 보기로 '전기(傳奇, 傳記, 前期, 前記)의 첫 모음 'ㅓ'는 짧고, '전기(電氣, 電機,)'의 첫 모음은 길다. 이런 모음의 길이는 규칙에 의해 예측 가능한 것이 아니고 각 낱말의 고유한 어휘적 특성의 하나이다.

2) ㄷ-덧나기: 합성어나 파생어를 이루는 낱말들 가운데 뒷말의 첫소리가 'ㅂ, ㄷ, ㄱ, ㅅ, ㅈ'와 같은 여린소리이면, 말의 표현을 똑똑히 하여 뜻을 분명히 밝히기 위해 두 말 사이에 /ㄷ/ 소리가 덧나는 일이 있다. 이때 앞 말이 모음으로 끝나는 고유어이면 사이시옷이 철자에 반영된다 (보기: 나룻배, 깃발, 고깃덩이). 그러나 한자말들끼리 결합하거나(보기:내과 內科[내꽈], 소장 訴狀 [소짱]), 또는 앞 말이 받침을 갖고 있으면(보기:손바닥[손빠닥]) 사이시옷이 표기되지 않는다. 사이시옷이 표기되지 않은 낱말들의 된소리 발음은 변동 규칙에 의해 자동으로 생성될 수 없다. 따라서 /ㄷ/-덧나기도 각 낱말의 고유한 어휘적 특성이다[1].

[1] 어간이 콧소리로 끝나는 동사, 형용사: 콧소리로 끝나는 동사와 형용사의 어간이 여린소리로 시작하는 씨끝과 결합하는 경우 어미의 여린소리는 된소리가 된다(보기:감다[감따], 감고[감꼬], 감지[감찌]). 이것 역시

3) ㄴ-덧나기: 합성어나 파생어를 이루는 낱말들 가운데 앞 말이 받침으로 끝나고 뒷말의 첫소리가 /ㅣ/ 또는 /ㅣ/-계열 이중모음으로 시작하면 말의 표현을 똑똑히 하여 뜻을 분명히 밝히기 위해 두 말 사이에 /ㄴ/이 덧나는 일이 있다(보기:늦여름[는녀름], 담요[담뇨], 나뭇잎[나문닢]). 그러나 위의 조건을 충족한 모든 낱말에 /ㄴ/ 소리가 덧나는 것이 아니므로(보기:전염[저념]) 이 현상을 규칙으로 처리할 수 없다.

4) 예외 발음을 갖는 낱말: '납량(納凉)'의 표준 발음은 예외적으로 [남냥]이 아니라 [나뱡]이다. 또 '맏형'도 단순히 변동 규칙을 적용하면 발음이 [마청]으로 되어야 하나 실제로는 [마텽]으로 발음된다. 이런 낱말들의 발음은 규칙에 의해 처리될 수 없으므로 불규칙 발음 사전에 등록되어야 한다.[2]

5) 다른 발음을 갖는 동일 철자의 낱말: 같은 철자를 갖는 낱말이 서로 다르게 발음되는 경우가 있다(보기:잠자리 [잠자리] '곤충의 한 종류', [잠짜리] '잠을 자는 곳'/ 아귀 [아귀] '물건의 갈라진 곳, 또는 아귀과의 바닷물고기', [아:귀] '굶어 죽은 귀신'). 이런 낱말을 올바른 발음으로 합성하기 위해서는 의미에 대한 판단이 필수적이다. 따라서 두 개 이상의 발음을 갖는 낱말들은 불규칙 발음 사전에 등록하여 관리해야 한다.

2. 한국어 불규칙 발음 사전 구축 방법

'표준 한국어 발음 대사전(이하 '표준 발음 사전')'에는 65,000여 개의 올림말이 실려 있다. 이들 올림말 가운데 '사랑[사랑]'과 같이 철자와 발음이 동일한 낱말들은 불규칙 발음 사전에서 제거되어야 한다. 이제부터 우리는 '표준 발음 사전'에서부터 철자와 발음이 동일한 낱말들을 효과적으로 정확하게 제거하는 방법을 논하고, 이어서 어떻게 불규칙한 발음을 갖고 있는 나머지 낱말들을 위에서 분류한 종류에 따라 구축할 수 있는가를 살펴볼 것이다.

1) 다른 발음을 갖는 동일 철자의 낱말 뽑기: 제일 먼저 '표준 발음 사전'에서 올림말의 철자가 같고 발음 표기가 서로 다른 낱말들을 뽑는다. 이런 낱말들에 대한 올바른 음성 합성을 위해서는 정확한 의미 분석이 필요하다. 아직 자동으로 낱말 의미의 애매성을 해결하는 기술이 개발되지는 못했으나 자연 언어 처리 기술의 발달에 대비하여 이런 낱말들의 목록을 확보하는 것은 중요하다.

철자와 발음이 동일한 낱말들을 뽑기에 앞서 이들 '동일 철자 - 다른 발음'을 갖는 낱말들을 먼저 뽑는 까닭은 '잠자리 [잠자리]/[잠짜리]'에서와 같이 두 발음 가운데 하나가 철자와 발음이 같은 경우가 많기 때문이다. 만약 철자와 발음이 같은 낱말부터 뽑아 제거한다면 '동일 철자 - 다른 발음'을 갖는 모든 낱말들을 제대로 뽑을 수 없다. 따라서 두 개 이상의 발음을 갖는 동일 철자의 낱말을 반드시 먼저 뽑아 특별히 관리해야 한다.

2) 철자와 발음이 일치하는 낱말 지우기: '표준 발음 사전'에서 '동일 철자 - 다른 발음'을 갖는 모든 낱말들을 제거한 다음, 나머지 올림말들에 '문자-음소 변환기[GTP; Grapheme to Phoneme]'를 돌려 잠정적인 발음을 생성한다. 이렇게 새로 만들어진 발음과 '표준 발음 사전'에 있는 기존의 발음을 비교하여 두 발음이 동일한 낱말들을 뽑아 제거한다. 이런 과정으로 불규칙한 발음을 갖는 낱말들만을 가려낼 수 있다.

일종의 ㄷ-덧나기이나 규칙적인 현상이므로 다음과 같은 변동 규칙을 세워 처리할 수 있을 것이다.
/ㄱ,ㄷ,ㅈ/ --> /ㄲ,ㄸ,ㅉ/ / [콧소리]_____

[2] '맛있다[마시따/마디따]나 '금융[그뮹/금늉]과 같은 낱말들은 두 발음 모두 표준 발음으로 인정된다. 이와 같이 복수 표준 발음을 갖는 낱말들은 음성 인식을 위해서는 불규칙 발음 사전에 등록되어야 하나 음성 합성에서는 이 가운데 하나만을 골라 합성해도 상관이 없으므로 불규칙 발음 사전에 포함하지 않아도 된다.

3) 장모음을 갖는 낱말 뽑기: 불규칙한 발음을 갖는 낱말들의 집합에서 발음에 장음 기호가 들어 있는 낱말들을 뽑아 하나의 목록을 만든다.

4) ㄷ-덧나기 낱말 뽑기: 불규칙한 발음을 갖는 낱말에서 장모음을 갖는 낱말들을 제거한 나머지 낱말들의 집합에서 발음에 된소리 /ㄲ,ㄸ,ㅃ,ㅆ,ㅉ/가 들어 있는 낱말들을 뽑아 하나의 목록을 만든다. 발음에 된소리가 있는 낱말들 가운데 '아까[아까]'나 '마땅[마땅]'과 같이 발음이 철자와 같은 낱말이나 '학교[학교]', 박쥐[박쮜]'와 같이 장애음 뒤에서 여린소리가 된소리로 되는 낱말들은 위의 2) 작업에서 모두 제거되었으므로 아직도 된소리 발음을 갖고 있는 낱말들은 예외 발음을 갖는 것들이다. 또 '헌법[헌:뻡]'과 같이 장모음을 갖는 동시에 ㄷ-덧나기를 하는 낱말들은 이미 장모음-불규칙 목록에 들어 있으므로 불규칙 발음을 찾는 데에는 아무런 문제가 없다.

5) ㄴ-덧나기 낱말 뽑기: 불규칙한 발음을 갖는 낱말들의 집합에서 발음에 /냐,녀,뇨,뉴,니/가 들어 있는 낱말들을 뽑아 하나의 목록을 만든다. 발음에 /냐,녀,뇨,뉴,니/가 있는 낱말들 가운데 '당뇨[당뇨]'와 같이 발음이 철자와 같은 낱말이나 '영장류[영장뉴]' 같이 변동 규칙에 의해 /냐,녀,뇨,뉴,니/ 소리가 나오는 낱말들은 위의 2) 작업에서 모두 제거되었으므로 아직도 /냐,녀,뇨,뉴,니/ 소리를 갖고 있는 낱말들은 예외 발음을 갖는 것들이다. 또 '방뇨[방:뇨]', '입장료[입짱뇨]'과 같이 장모음을 갖거나 ㄷ-덧나기를 갖는 동시에 ㄴ-덧나기를 하는 낱말들은 이미 장모음-불규칙 이나 ㄷ-덧나기 목록에 들어 있으므로 불규칙 발음을 찾는 데에는 아무런 문제가 없다.

6) 예외 발음을 갖는 낱말 뽑기: 이런 과정을 거친 뒤에 남는 낱말들은 모두 '납량(納凉) [나뱡]', 또는 '맏형[마텽]'과 같이 예외 발음을 갖는 목록이 된다.

불규칙 발음 사전을 만들 때 한 가지 주의할 것은 어간 마지막에 오는 받침의 처리이다. 한국어에서 다른 장애음 앞이나 어말에 오는 자음들은 '일곱 끝소리 되기' 규칙에 의해 /ㄲ, ㅋ/는 /ㄱ/으로, /ㅍ/는 /ㅂ/으로, /ㅅ, ㅈ, ㅊ, ㅌ/는 /ㄷ/으로 바뀐다(보기: 꽃과[꼳과], 부엌도[부억또], 옷[옫]). 그러나 불규칙 발음 사전에서 어말에 오는 받침들은 원래 자음을 그대로 유지해야 한다. 그 까닭은 이 소리들이 중화한 대표음으로 표기되면 '소리의 이음 규칙(연음 규칙)' 이 적용될 때 잘못된 발음이 만들어지기 때문이다. 예를 들어 '나뭇잎'의 발음을 일곱 끝소리 되기 규칙에 의해 [나문닙]으로 적는다면 그 뒤에 주격 조사 '-이'가 연결되었을 때, [나문니비]로 잘못된 발음이 도출된다. 이런 잘못된 결과를 피하기 위해서는 불규칙 발음 사전에 '나뭇잎'의 발음을 [나문닢]으로 적어야 한다. 그리고 명사와 조사, 용언의 어간과 어미의 결합이 다 끝난 뒤에 어말에 오는 자음들에 일곱 끝소리 되기 규칙을 적용해야 한다.

3. 불규칙 발음 사전의 이용 방법

음성 합성에 불규칙 발음 사전을 적용하는 방법은 정확한 형태소 분석기가 개발되어 있느냐 또는 아니냐에 따라 다르다.

형태소 분석기가 구현되어 있는 경우, 대상 어절을 실사(명사나 동사, 형용사의 어간) 부분과 허사(조사나 어미) 부분으로 분리한 다음, 실사를 불규칙 발음 사전에서 검색한다. 만일 해당하는 낱말이 불규칙 발음 사전에서 검색되면 이 실사의 철자를 발음으로 대치한 뒤 모든 어절에 대해 문자-음소 변환기를 돌려 바라는 발음을 얻어 낸다.

반대로 형태소 분석기가 없는 경우에는 대상 어절을 왼쪽에서 오른쪽으로 비교해 가며 최장일치가 되는 문자열을 찾아 이를 실사 부분으로 보고 철자를 발음으로 치환한 뒤 모든 어절에 대해

문자-음소 변환기를 돌린다.

이런 과정이 끝난 다음, 음소-변이음 변환기를 돌려 얻어진 음성 기호에 대응하는 음성 데이터를 찾아 기계로 소리를 합성하면 음성 합성이 완성된다.

4. 끝맺는 말

이 연구에서 우리는 '표준 한국어 발음 대사전(1993 한국방송공사 편찬)'을 바탕으로 불규칙 발음 사전을 구축하는 방법에 대해 자세히 살펴보고 이어서 이를 이용하여 한국어를 음성 합성하는 방법을 논의했다. 이와 같은 한국어 불규칙 발음 사전의 구축은 보다 정확한 발음의 한국어 음성 합성을 가능하게 한다는 점에서 의의가 있다. 그러나 아직 자동으로 낱말 의미의 애매성을 해결하는 기술이 개발되지 않아 '잠자리 [잠자리]/[잠짜리]'와 같이 '동일 철자 - 다른 발음'을 갖는 낱말에 대해 해결할 수 없는 한계를 갖고 있다. 자연 언어 처리 기술이 발달해 이 문제까지도 해결한다면 컴퓨터를 비롯한 전자 기기와 우리말로 정보를 주고받을 날이 올 것이다.

끝으로 음성 인식에 있어서는 불규칙 발음 사전의 효용 가치가 음성 합성의 경우보다 많이 떨어진다. 음성 인식은 음파를 변이음 기호로 바꾸고 이를 다시 음소로 변환한 뒤, 마지막으로 음소를 철자로 바꾸게 된다. 이런 작업을 위해서는 모든 가능한 발음과 그에 대응하는 철자를 갖고 있는 일반 '발음-철자' 사전이 더 유용하다. 만약 불규칙 발음을 갖는 낱말만 발음 사전에서 찾고 나머지 낱말들을 규칙에 의해 말음에서 철자로 바꿔가는 방법을 취한다면 음성 인식은 매우 복잡한 과정을 거쳐야 할 것이다. 이런 이유로 불규칙 발음 사전은 음성 인식보다는 음성 합성에 더 유용하다.

참고문헌

김기호 (1994) '음성 합성과 음성 인식에 있어서의 음성학과 음운론의 역할', 음성학 학술 대회 자료집, 대한 음성학회.

김선철 (1990) '현대 국어 음운 규칙의 순서 매김 원리에 대하여', 언어학 연구 5, 서울대학교 대학원 언어학과.

유재원 (1994) '연속 음성 인식을 위한 음성 단위 발음 사전 구성 방법 연구', 전자통신연구소 위탁 연구 과제 보고서.

한국 방송 공사 (1993) '표준 한국어 발음 대사전', 어문각.

허웅 (1988) '국어 음운학', 샘문화사.

____ (1984) '국어학', 샘문화사.