한국어 리듬단위 생성기 연구

-낭독체 중심으로-

정회선(서울대 언어학과)

〈차례〉

1. 머리말

음성합성은 최근에 들어 크게 주목받고 있는 언어처리 기술이다. 음성을 합성을 한다는 것 자체가 관심거리였던 시기는 지났고 지금은 그 자연성을 높이기 위한 여러 방안들이 연구되고 있다. 이러한 연구들 중 특히 억양에 관한 연구는 상당히 활발하게 이루어지고 있고 여러 곳에서 그 구현을 시도하고 있는 것으로 보인다.

무제한 음성합성을 전제로 한다면 이러한 연구를 구현하기 위해서는 전제 조건이 필요하다. 그것은 과연 어느 곳에서 숨을 쉬고 어느 곳에서 끊을 것인가, 그리고 어떤 곳을 불여야 할 것인가가 자동적으로 결정되어야 한다는 것이다.

이러한 일을 수행할 수 있는 엔진을 제작하기 위해서는 당연히 한국어의 리듬에 관한 기초 연구들을 토대로 하여야 하고 현재의 문헌어 처리 기술을 고려한 엔진의 설계가 이루어져야 한다. 한국어 리듬 분야에 있어서 기반으로 삼아야 할 기존의 연구로는 이현복(1989), 이호영(1991), 성철재(1995) 등을 들 수 있다. 그리고 이를 구현하기 위한 노력으로는 김형욱(1992), 강용범(1993), 등을 들 수 있다.

그러나 위의 두 실험적인 노력으로 한국어 리듬을 구현하기 위한 모델이 완전히 이루어졌다고 보기는 곤란하다.

따라서 본 연구는 한국어 리듬단위 생성기를 염두에 두고 지금의 텍스트처리 기술

수준에서 추출 가능한 요소들을 고려하여 적용가능한 규칙과 구현 가능성에 대해 알아보고자 한다.

2. 본문

끊어읽기를 자동적으로 구현하기 위해서 고려해야 할 우선적인 대상은 붙여읽을 가능성이 높은 경우의 목록과 끊어읽을 가능성이 높은 경우의 목록을 작성하는 것이다. 이를 위해 방송뉴스를 녹취하여 분석하였다. 이를 토대로 말토막 하나에 들어가는 음절 수와 어절 수를 통계적으로 평균을 내는 한편 문법형태소들 사이의 끊어읽기 위계를 정한다. 한 문장 안에 나타나는 문법형태소들이 가지고 있는 디폴트 값은 문맥에 따라 조정된다.

2.1 기초 자료 조사

2.1.1자료수집 과정
실제 발화의 녹취를 위해서 다음과 같은 과정을 밟았다.

1) KBS 라디오 뉴스 녹음
   분량: 약 95분
2) 뉴스 내용 녹취
3) 8명에게 녹음된 내용을 들려주어 휴지를 표시하게 함.
4) 자료입력

2.1.2 휴지 경향

휴지등급	빈도	비율	누적빈도	누적비율
0	4798	53.43	4798	53.43
1	1104	12.29	5902	65.72
2	598	6.66	6500	72.38
3	395	4.40	6895	76.78
4	311	3.46	7206	80.24
5	330	3.67	7536	83.92
6	355	3.95	7891	87.87
7	551	6.14	8442	94.01
8	538	5.99	8980	100

2.1.3. 말마디당 평균 음절 수 및 어절 수

이들 총 8980 단어에 대한 분석을 한 결과 한 어절의 평균 음절 수는 2.623인데 반해 말마디 하나에 들어가는 평균 음절 수는 13.277이며 평균 5.062 개의 어절이 하나의 말마디를 이룬다.(편의상 5명 이상이 휴지가 있다고 생각한 곳을 말마디의 경계가 있다고 보기로 함.)

##총 8980 WORDS에 대한 분석 결과
###어절당 평균 음절 수: 2.623
###말토막당 평균 음절 수: 13.277
###말토막당 평균 어절 수: 5.062

2.1.4. 형태소의 끊어읽기 위계

우리말은 어절과 어절 사이를 연결해 주는 것이 대부분 조사와 어미이다. 그밖에 관형사와 부사, 감탄사 등이 있으나 이들은 문장의 발화시 큰 변이가 없으므로, 우리말에 나타나는 조사, 어미의 빈도수를 고려하는 것이 어절 간의 경계를 찾는 데 중요한 항목이 된다. 다음은 녹취 자료에 나타난 조사와 어미의 목록이다.

* 자료 NEWS.KHS에 나타나는 조사, 어미 및 특이 어휘 목록

-가	-된	-어서	-이	뒤	중
-거나	-로	-에	-이며	등	쯤
-고	-를	-에게	-인	따라	채
-과	-며는	-에는	-챈	따라서	한편
-까지	-면	-에서	-쳐	또	할
-나(서)	-면서	-온	-한	마는	함께
-는	-부터	-와	-해	분	후
-던	-써	-은	결과	씨	
-데	-아	-을	경우	오늘	
-도	-어	-의	도록	있어	

문법형태소들은 끊어 읽기에 있어서 일정한 위계를 지니고 있다. 이 위계는 통계자료를 통해 어느 정도 추정할 수 있다. 위에 제시한 뉴스 녹취자료의 분석 통계를 통해 이들의 기본적인 위계를 추정해 보고자 한다. 문법형태소들을 휴지/빈도수 비율(이하 휴지지수) 순으로 나열해 보면 다음과 같다. (단 영형태소는 제외하였다.)

-위계목록:sample.dat

어말목록	품사	휴지/빈도	휴지지수
-의	조사	14/369	0.037
-을	조사	37/427	0.086
-를	조사	27/278	0.097
-에	조사	66/377	0.175
-와	조사	19/105	0.180
-이	조사	61/334	0.182
-도록	어미	5/25	0.200
-가	조사	47/176	0.267
-보다	조사	5/18	0.277
-ㄴ/는	어미	193/606	0.318
-고	어미	133/413	0.322
-에게	조사	8/24	0.333
-에서	조사	65/192	0.338
-부터	어미	20/58	0.344
-과	조사	42/122	0.344
-는	어미	131/301	0.435
-까지	조사	18/40	0.450
-로	조사	59/123	0.479
-은	조사	144/283	0.508
-는	조사	130/249	0.522
-해	어미	61/99	0.616
-아	어미	11/16	0.687
-돼	어미	7/10	0.700
-어	어미	19/27	0.703
-자	어미	5/7	0.714
-나	어미	28/38	0.736
-도	조사	28/37	0.756
-나	조사	19/25	0.760
-서	어미	119/148	0.804
-가	어미	12/14	0.857
-며	어미	27/27	1.000

위 목록을 보면 '-의'가 가장 휴지지수가 낮다. 그리고 어미는 조사보다 휴지 지수가 높다. 이런 어미 중에서도 종결어미는 문장의 끝에 오므로 가장 높은 위계에 속하는 당연하며, 접속어미는 연결어미 보다 위계가 높다. 이것을 도식화 하면 다음과 같다.

종결어미 > 접속어미 > 연결어미 > 보조사 > 격조사

그러나, 어미와 조사의 위계에서 주의할 점은 이것이 환경을 고려하지 않은 디폴트 값이라는 것이다. 즉, 연결어미 뒤에 보조동사나 다른 동사가 바로 연결될 경우에는 오히려 붙여읽기 요소에 속하게 되므로, 위의 위계에서 고려되지 않는다.

또한 어미 중에서 관형형 어미는 조사 일반보다 위계가 낮다. 그러나 이것도 다른 요소를 고려해야 한다. 즉, 단순한 수식의 경우에는 위의 위계가 맞지만 이것이 절로 쓰일 경우나 그 뒤에 명사가 바로 연결되지 않고 다른 수식어가 올 경우에는 쉼의 정도가 보다 커진다.

어절 중에는 문법형태소가 붙지 않은 독립어들이 있다. 관형사, 부사, 감탄사 그리고 접속사가 그것이다. 이들 중 관형사를 제외한 다른 품사들은 독립적인 말토막을 형성할 수 있다. 또한 용언이나 체언에 영형태소가 연결된 경우도 있는데, 이들은 뒤에 오는 요소와 밀접히 관련되는 것이 보통이다.

이를 형태소 위계와 연결하여 도식화하면 다음과 같다.

종결어미 > 접속어미 > 부사/감탄사/접속사> 연결어미
> 보조사 > 격조사 > 관형사 > 실사+영형태소

2.1.5. 붙여읽기요소

우리말의 경우에 어절단위로 띄어쓰기 때문에 휴지요소의 발견에 있어서 끊어읽기보다는 붙여읽기요소를 추출 하는 것이 더욱더 중요하다. '-ㄹ 수'나 '-ㄴ 지' 등이 그것이다. 이와 더불어 다음과 같은 요소들은 앞에 명사나 격조사가 올 경우에 붙여읽는 것이 보통이다. 편의상 이들을 후치사라고 하겠다. 또한 불완전명사의 경우에도 동일한 현상이 나타난다. 휴지요소 발견에 있어서 이들은 모든 요소에 우선한다. 이에 명사와 명사의 연결을 추가하여 가설로 삼는다. 이를 다시 도식으로 나타내면 다음과 같다.

가설 ㄱ)

종결어미 > 접속어미 > 부사/감탄사/접속사> 연결어미 > 보조사
> 격조사 > 관형사 > 명사+명사> 실사+영형태소 > 붙여읽기요소

-후치사 목록 (일부)

가운데/간/같이/거치여/걸쳐/걸쳐서/걸친/곁/고사하고/관계/관하여/관한/관해/까지/나머지 /내/너머/대로/대하여/대한/대해/더불어/더불어/동안/뒤/때문에/또한

위에/위로/위하여/위한/위해/더불어/으로써/의하여/의한/의해/있어/있어/있어서/중/지나서/진/채로/처럼/처하여/처한/처해/통/통하여/통한/통해/하/하여금/하여금/한테/함께/향하여/향한/향해/후

-불완전명사 목록 (일부)
/것/결/개/거리/게/겸/경/고/그루/길/김/깃/까닭/나름/나위/나절/내/내기/녀석/녘/또래/류/리/마리/마음/만/만큼/말/망/망정/모/모양/무렵/물/바/바람/바리/밖/발/번/벌/법/부/분/빨/뿐/사/사람/산/상/새/새록/생/서슬/설/섬/성/세/셈/손/수/수록/시/양/어/

2.2. 문법요소 연쇄를 고려한 수정

위 자료조사의 의미는 일반적 경향과 불변요소의 경험적 발견에 있다. 이러한 발견위에 문법관계에 따른 요소들을 보완하여 전체적인 체계를 수립한다. 이를 위해 다음과 같은 가설을 추가로 설정하여 위의 가설 ㄱ)의 상위 규정으로 삼는다.

가설 ㄴ. 직접 수식관계인 어절들의 인접 연쇄는 붙여읽을 가능성이 높다.
가설 ㄷ. 떨어져 있는 어절을 수식할 경우 끊어읽을 가능성이 높다.

2.2.1
위의 가설을 다음의 짧은 문장들의 예를 통해 검토해 보자.

1) 아름다운 1 꽃
   작고 2 아름다운 1 꽃

2) 아름다운 1 꽃
   매우 1 아름답고 2 향기로운 1 꽃

2) 저녁을 1 먹어야겠다.
   저녁을 2 어서 1 먹어야겠다.
   저녁을 2 어서 3 배부르게 1 먹어야겠다.

3) 빨리 1 가라.
   빨리 2 집으로 1 가라.
   빨리 2 집으로 2 달려 1 가라.
   빨리 3 철수의 2 집으로 3 달려 1 가라.

(숫자는 한 문장내에서 끊어읽기의 상대 등급. 작을수록 붙일 가능성이 높다.)

위의 가설을 다음의 긴 문장을 통해 살펴보자.

4) 바람과 2 햇님이 3 서로 3 힘이 3 더 2 세다고 다투고 1 있을 1 때 3 한 2 나그네가 3 따뜻한 2 외투를 2 입고 걸어 1 왔습니다.

5) 그들은 3 누구든지 3 나그네의 2 외투를 3 먼저 2 벗기는 1 이가 3 힘이 3 더 2 세다고 2 하기로 2 결정했습니다.

위의 자료를 통해 자연스런 문법 연상관계가 이어지는 경우 우선적으로 붙여 읽어 어색하지 않음을 알 수 있다. 여기서 자연스런 문법 연상관계를 정의할 필요가 있다. 적어도 아래의 예들은 자연스런 문법연상관계인 것으로 보인다.

ㄱ) 용언[관형형] 명사
ㄴ) 부사 용언
ㄷ) 부사 용언[관형형] 명사
ㄹ) 부사어 동사
ㅁ) 목적어 타동사
ㅂ) 체언[주격] 자동사

단 ㄴ)과 ㄷ)의 경우 용언에 따라 앞의 부사는 그 용언을 수식할 수 있는 부사이어야 한다. 즉 적어도 문장부사의 경우는 제외된다. 또한 "빨리 3 먼 1 곳으로 2 떠나라!"와 같은 문장에서 '빨리'와 '먼'은 부사 용언의 연쇄이기는 하나 빨리가 형용사인 '먼'을 수식할 수 없다. 그런데 이 정보는 렉시콘에 미리 지정할 수 있으므로 쉽게 "아주 2 먼 1 곳으로 2 떠나라!"와 구별할 수 있다.

3. 결론

위의 자료와 분석 및 수정을 통해 우리는 우리말 리듬단위를 자동으로 생성함에 있어서 높은 수준의(상대적으로 확률이 낮고 시간이 많이 걸림) 구문 분석기를 사용하지 않아도 문법형태소 고유의 정보와 인접관계들을 고려함으로써 확률 높은 결과물에 도달할 가능성을 타진해 보았다. 이 연구를 완결시키기 위해서는 여러 문법형태소들의 인접을 더 폭 넓게 연구하는 한편, 그들 사이의 위계를 결정하고 그러한 요소들이 복합적으로 작용할 때 적용될 결합규칙을 찾아야 할 것이다.

<참고문헌>

강용범(1993) 무제한 음성합성 시스템을 위한 문장구조 추출에 관한 연구. 「 제 10회 음성통신 및 신호처리 워크샵 논문집 」 한국 음향학회.

김형욱(1992) 한국어 문장-음성 변화기의 운율 제어용 구문분석기 「 제 9회 음성통신 및 신호처리 워크샵 논문집 」 한국 음향학회.

성철재(1995) 「 한국어 리듬의 실험음성학적 연구 -시간 구조와 관련하여 」 서울대학교 박사학위 논문.

이현복(1986). 한국어 음성의 합성과 인식에 관한 음성언어학적 고찰. 「 한글 」 194. 한글학회. 55-72.

———(1989) 「 한국어의 표준발음 」 교육과학사. 개정판(1993).

———(1995). 「 음성합성을 위한 운율생성모델 제어규칙 연구 」. 한국전자통신 연구소.

이호영(1991) 한국어의 리듬. 「 한국어 연구논문 」 제 28집. KBS 한국어 연구회.