구문분석을 이용한 한국어 음성합성의 운율생성 연구

A study on the prosody generation in Korean speech synthesis using sentence structure analysis

백 승 권, 김 원 철, 한 민 수

(Seung-Kwon Beack, Won-Cheol Kim, Minsoo Hahn)

한국정보통신대학원대학교 공학부

E-mail : skbeack@icu.ac.kr

ABSTRACT

In this paper, we presented the prosody analysis results of five selected words according to its usage in a sentence, i.e., the part of sentence (PoS) while changing the type of sentences such as simple, conjugate, and complex sentences. The selected five Korean words were "U-Ri-Na-Ra", "Bul-Kuk-Sa", "Uh-Muh-Ni", "Han-Ra-San", and "Gang-A-Ji". These five words were used as a subjective, an objective, and an adverb in each simple, conjugate, and complex sentence. The pitch, energy, and duration of each word were then analyzed and used for the synthetic speech prosody improvement. The subjective test on the prosody improvement showed that more than 50% of our listeners are affirmative to the prosody improvement of the synthetic speech.

I. 서 론

음성합성 기술은 인간과 컴퓨터 사이의 가장 자연스런 의사 전달 형태인 음성언어를 이용한 man-machine interface 중 음성 출력 기능을 담당하는 중요한 기술 중의 하나이다. 따라서 합성음의 자연성 및 명료도로 나타내어지는 품질은 음성 인터페이스의 출력단의 품질을 결정하게 되며 이는 사용자가 음성 인터페이스를 이용할 것인지 아닌지를 결정하는 중요한 요소가 된다.

현재 국내에서는 산업체, 연구소, 학계 등에서 낭독체 한국어 합성기를 보유하고 있으며 품질도 선진국에 비하여 큰 차이를 느낄 수 없는 수준이다. 그러나 세계적으로 음성 합성기의 성능이 명료도에서는 어느 정도 합격점을 받고 있으나 자연성 측면에서는 아직 합격점을 받기에는 요원한 수준이며 국내의 한국어 합성기도 비슷한 수준에 머물고 있다. 즉, 비교적 단순한 문장을 상대적으로 단순화된 운률 규칙을 이용하여 합성하고 있는 수준인 것이다.

따라서 본 연구에서는 동일 단어의 한국어 문장 구성 성분 (PoS)에 따른 운률 특성 변화를 분석하고 이를 이용하여 보다 자연스런 합성음의 운률을 구현하려는 것이다. 이는 향후 한국어에 대한 자연어 처리 기술이 텍스트 만을 적용 대상으로 하는 것이 아니라 한국어 음성합성기에도 적용되어 합성음의 운률을 개선하고 따라서 사용자에게 보다 자연스런 합성음을 들려주기 위한 출발점이 될 것이다.

II. PoS(Part of Sentence) 추출

2.1 낭독체 음성 DB 구축

동일 단어의 PoS 별 변화에 따른 운률 특성 변화를 조사하기 위한 낭독체 음성 DB 구축을 위하여 사용된 단어는 '우리나라', '불국사', '한라산', '어머니', '강아지'의 5 단어이다. 이 5 단어의 선택은 일단 발성자에게 익숙하여 자연스럽게 발성할 수 있는 단어 중에서 선택되었다. 한편, 모음 만으로 이루어진 단어는 지속시간 및 pitch 의 변화가 상대적으로 크며 자음이 많은 단어는 지속시간 및 pitch 의 변화가 상대적으로 작다는 일반적인 특성이 있으므로 한 단어 내의 자음의 수가 비교적 다양하게 존재하도록 하기 위하여 선택되었다. 또한 선택 단어는 운률 변화가 보다 심할 것으로 예상되는 감성적인 단어인 '우리나라', '어머니', '강아지'와 그렇지 않은 단어인 '한라산', '불국사'로 구성되었음을 알 수 있다.

사용된 문장은 단문, 중문, 복문의 3 가지였으며 각 문장 종류에 대해 동일 단어가 주어, 목적어, 부사어(관형어)로 쓰이도록 하였으며 해당 단어 이외의 부분은 가능한 한 동일 단어들을 사용하여 구성하였다. 이 문장을 이용하여 낭독체 음성 DB 를 구성하기 위해 표준말을 사용하는 서

울에서 성장한 20 대 중반 남성화자 3 인과 여성화자 3 인이 curtain 으로 둘러싸인 조용한 사무실에서 각각 1 회씩 발성한 문장을 10 kHz 로 샘플링하고 16 비트로 양자화하여 저장하였다. 결과적으로 총 270 개의 문장 DB 와 단어 DB 를 구축하였다.

본 연구에서는 PoS 를 추출할 때 구간정보의 정확도를 높이기 위하여 전문가 2 인이 디지털 파형편집기를 이용하여 문장 중에서의 PoS 를 수동으로 추출하였다.

2.2 PoS 운율정보 추출 방법

사람 말소리의 운률정보는 음의 높낮이를 나타나는 피치, 음의 세기를 의미하는 에너지, 그리고 음의 길이로 나타나는 지속시간 정보로 구성되며 합성음의 자연성은 합성음의 운률, 즉 상기 피치, 에너지 및 지속시간의 특성이 얼마나 native speaker 와 비슷하냐에 의하여 결정된다 [1]. 따라서 한국어 합성음의 자연성을 향상시키기 위해서는 한국어 운률 정보의 분석이 매우 중요하다.

피치검출은 그 응용분야가 음성 분석, 합성, 인식 및 코딩 등 음성처리 기술 분야 전반에 걸쳐 무척 다양하므로 현재까지 많은 연구가 활발히 진행되어 왔다 [2][3]. 한편 피치 정보는 피치동기식 특성 분석 뿐만 아니라 음성신호를 유성음/무성음/묵음 구간으로 분리하고 각 구간에 따른 합성음원의 변화 정보나 코딩 기법의 변화 등에도 유용하게 이용되어 왔다 [4][5].

피치 검출법은 일반적으로 시간영역법, 주파수영역법, 그리고 시간-주파수 혼성 영역법으로 분류될 수 있다. 본 연구에서의 음성 데이터는 비교적 잡음이 없는 환경에서 수집되었으므로, 이에 대한 피치를 검출하는 알고리즘으로 시간영역법이 효율과 성능면에서 가장 적합하다고 판단하여 시간영역 피치검출법을 사용하였다. 사용된 피치 검출법은 AMDF(Average Magnitude Difference Function)을 사용하였으며, 식 (1)에 나타내었다 [6].

$$r_n(k)=\sum_{m=-\infty}^{\infty}\left|x(n+m)w_1(m)-x(n+m-k)w_2(m-k)\right| \quad (1)$$

$w(n)$: 창함수, $x(n)$: 음성데이터

에너지 정보는 정규화된 단구간 에너지를 10 msec 간격으로 다음의 식 (2)를 사용하여 구하였다.

$$E_n=\frac{1}{N}\sum_{m=-\infty}^{\infty}\left[x(m)w(n-m)\right]^2 \quad (2)$$

$w(n)$: 창함수, $x(n)$: 음성데이터이

지속시간 정보는 높은 정확성이 요구되기 때문에, PoS 별 대상 단어의 지속시간정보를 추출할 때에는 정확도를 최대로 높이기 위하여 전문가 2 인이 디지털 파형편집기를 이용하여 검출하고 들어서 확인하는 과정을 반복하여 수동으로 작업하였다.

III. PoS 추출 결과 및 분석

3.1 Pitch 추출결과 및 분석

본 연구에서는 수집된 남녀 각각 3 인의 음성 DB 에 대하여 pitch 정보를 분석하여 남녀에 대한 일반적인 pitch 특성 및 각 단어의 문장성분과 종류에 따른 pitch contour 변화를 살펴보았다.

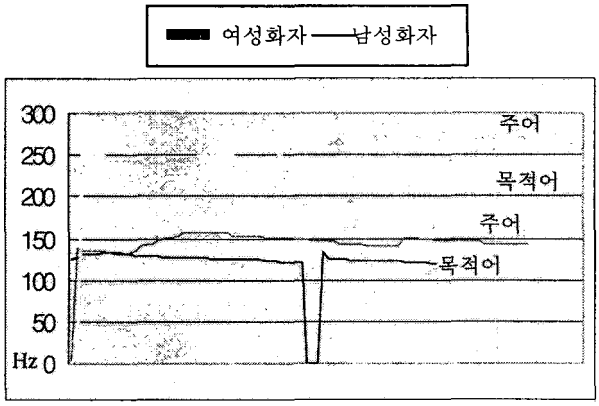

그림 1. Pitch contour,'우리나라': 단문

그림 1 은 '우리나라'의 단문에서 문장성분의 변화에 따른 pitch contour 를 나타낸다. 그림에서 보면 남성, 여성화자 모든 경우에 대해 주어인 경우가 목적어일 때 보다 다소 높은 pitch contour 를 가짐을 알 수 있다. 그림 1 은 분석된 결과 중 일반적인 사례를 예를 보여주는 것으로써 거의 모든 경우에 주어로 쓰인 경우가 목적어나 부사어 (또는 관형어)로 쓰인 경우보다 pitch 값이 다소 높게 나타난다. 그러나 문장성분의 변화에 따른 pitch contour 는 큰 변화를 보이지 않았다.

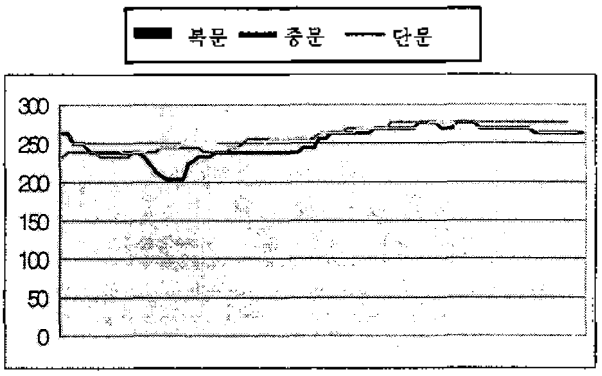

그림 2. 문장종류에 따른 pitch contour, '우리나라' 주어

그림 2 는 동일 단어, 즉 '우리나라'가 같은 문장 성분, 즉 주어로써 사용되었으나 사용된 문장의 종류를 단문, 중문, 복문으로 달리할 때 pitch contour 가 어떻게 변하였는지를 나타낸다. 그림 2 에서 알 수 있는 사실은 문장종

류의 변화에 따른 pitch의 변화는 무시할 만하다는 것이었다. 따라서 pitch contour는 문장성분이 바뀔 때 다소 변화하지만 문장의 종류에 따라선 거의 변화하지 않는다고 말할 수 있다.

3.2 Energy 추출결과 및 분석

Energy는 사람의 말소리와 운율 정보 중 말의 세기를 나타낸다. 본 연구에서는 먼저 문장성분의 변화에 따른 동일 단어의 energy contour를 분석하고 그 결과를 바탕으로 energy contour를 조절하는 규칙을 기존의 합성기에 적용함으로써 개선된 합성 음질을 제공하고자 하였다.

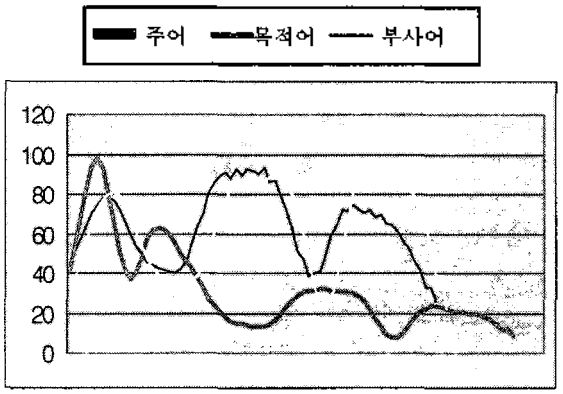

그림 3. '우리나라', 중문에서의 energy contour

그림 3은 '우리나라'에 대한 중문에서의 문장성분에 따른 energy contour를 보였다. 주어로 쓰일 경우 단어의 시작 부분에서 에너지 변화가 심하며 단어 끝으로 갈수록 변화량이 감소하면서 평탄하게 변함을 볼 수 있다. 부사어나 목적어로 쓰일 경우의 energy contour는 거의 동일한 특성을 보인다. 즉, 특정구간에 대해 큰 변화를 보이지는 않지만 전체적인 energy 변화량은 주어일 때보다 심하다. 이러한 energy contour 결과를 얻게된 것은 일반적으로 주어 부분이 문장의 핵심 단어가 될 경우 이를 강하게 발성하기 때문이라 판단된다. 목적어나 부사어로 쓰인 경우 해당 문장에서 문맥상 핵심단어로 쓰일 때에는 강조하여 발성하려는 성향으로 에너지 변화가 커질 수 있다. 그러나 이 경우는 주어와는 다르게 단어의 어미 부분에서도 억양을 높여 발성하려는 성향으로 단어의 어미 부분에서도 상대적으로 큰 energy를 갖게 된다.

3.3 Duration 추출결과 및 분석

Energy의 경우와 마찬가지로 동일 단어의 문장 내에서의 구성성분 변화에 따른 지속시간의 변화를 살펴보고 또 문장의 종류가 바뀔 때의 변화를 분석하여 일반적인 변화 패턴을 추출하고자 한다. 또한 energy와 duration 사이의 상관관계도 고찰하였다.

Duration 정보를 energy 크기와 연관하여 음성 DB를 분석하는 이유는 발성 단어의 구간별 지속시간이 일반적으로 평균 에너지가 커지면 비례해서 커지는 경향이 있기 때문이다.

표 1. 중문에서'우리나라'의 구간별 평균 energy와 duration

문장성분 \ 경계		우리	나	라	Total
주어	Energy	102	38	27	56
	Duration	141msec	135msec	141msec	418msec
목적어	Energy	117	134	111	121
	Duration	88msec	145msec	114	349msec
부사어	Energy	137	108	125	122
	Duration	113msec	154msec	128msec	396msec

표 1과 같이 중문에서의 '우리나라' 단어 발성에 따른 구간별 energy와 duration 정보를 살펴보면 앞절에서 밝힌 바와 같이 주어로 사용될 경우 단어의 시작 부분이 강조됨에 따라 단어의 끝부분에서보다 큰 energy 값을 가짐을 관찰할 수 있다. 이 부분에 대한 duration 값을 구해보면 다른 구간의 음운에서의 duration 보다 길게 나타나고 있어서 energy가 커지면 지속시간이 길어진다는 일반적인 가설을 뒷받침해 줌을 알 수 있다. 그러나 전체적인 평균 energy 값을 보면 부사어나 목적어로 쓰였을 경우가 더 큼을 알 수 있다. 표 1에서 전체 단어에 대한 평균 energy와 duration을 구해 보면 부사어일 경우 평균 energy가 가장 크게 나타났으며 주어일 경우가 가장 작게 나타 났다. 반면에 duration은 주어일 경우에 더 길었다. 결국 duration과 energy는 서로 상관관계가 있지만 음절단위에 대한 상관관계가 크며 전체 단어에 대해선 적음을 알 수 있다.

표 2는 문장성분 및 문장종류에 따른 duration이 가장 길게 발성되는 성분에 대한 통계치이다. 표 2를 통해 알 수 있는 것은 각 문장성분에 대한 duration 정보는 단문일 때 가장 길게 나타나는 경우가 많다는 것이다. 그러나 중문과 복문에 대해선 편차가 크지 않으므로 이를 통해 일반적인 패턴을 추출하기는 어렵다고 판단된다.

표 2. 문장종류에 따른 최대 duration 변화치 갯수

문장종류 \ 문장성분	주어	목적어	부사어관형어
단문	9	7	8
중문	4	6	7
복문	7	7	5

IV. PoS 정보에따른 운율생성규칙 작성 및 성능평가

PoS 정보에 따른 운률 생성 규칙은 '불국사'에 대한 pitch, energy, duration 정보를 분석한 결과를 바탕으로 합성 문장 중 해당 단어의 energy 와 pitch 및 지속시간을 조정하였다.

성능평가는 '불국사' 한 단어에 대하여 새로운 운률 규칙을 적용한 합성음과 (즉 해당 단어 이외에는 기존의 합성 운률이 그대로 적용되었음) 기존의 문장단위 운률 패턴을 갖는 합성음을 생성하여 무작위로 남녀 각 5 인에게 들려주고 보다 자연스런 합성음을 선택하게 하였다. 그 결과를 표 3 에 요약하였다.

표 3 에서 보면 약 52 %가 새로운 운률 규칙이 적용된 합성음이 보다 자연스럽다고 선택하였으며 약 27%가 차이가 없다고 판정하였다. 이는 PoS 별 운률 패턴 생성 규칙이 합성음의 자연성 개선에 도움이 된다고 해석할 수 있다. 한편 단문의 경우 약 70 %가 새로운 합성음이 보다 자연스럽다고 선택하였으며 약 7%가 차이가 없다고 판정하였다.

표 3. 새 운률과 기존 운률에 따른 자연성 선호도 조사 결과 (단위: 명)

선호도 / PoS 문장종류	새 운률	기존 운률	차이 없음
단문 주어	6	3	1
단문 목적어	7	2	1
단문 부사어	8	2	0
중문 주어	5	4	1
중문 목적어	6	1	3
중문 부사어	4	1	5
복문 주어	7	1	2
복문 목적어	2	3	5
복문 부사어	2	2	6
Total	47	19	24

이는 단문의 경우 새로운 합성 운률에 의한 자연성 개선 효과는 더 크지만 개인별로 보다 선호하는 운률 패턴이 존재한다고 해석할 수 있다. 일반적으로 경북 지방에서 성장한 사람은 주어, 목적어, 부사어 구별없이 2 음절 명사의 경우 강세가 앞에 오는 것을 선호하는 한편 부산지방에서 성장한 사람은 강세가 뒤에 오는 것을, 서울 및 경기 지방에서 성장한 사람은 강세가 PoS 별로 약간 변하는 것을 선호한다고 알려져 있으며 이는 10 명의 피험자의 선호도 평가표를 분석해 봐도 대체적으로 일치함을 알 수 있었다.

한편 문장이 복잡해져 감에 따라 새로운 운률에 따른 자연성 개선 효과가 감소하였다. 이는 전체 문장 중 운률이 다른 부분이 해당단어 뿐이므로 문장이 길어질수록 변별력이 적어지는 것에 기인한다고 판단된다.

V. 결론

본 논문에서는 향후 자연어 처리기의 구문 분석 결과를 이용하여 합성음의 자연성을 개선할 수 있는 방안을 모색하기 위한 기본 연구로 동일 단어의 PoS 별, 문장 형태별 운률 특성을 분석하고 이를 이용하여 새로운 운률 생성 규칙을 작성한 후, 이를 적용한 합성음과 기존의 문장 단위별 운률 생성 규칙을 적용한 합성음을 비교 청취하여 자연스러운 것을 선택하도록 하고 그 결과를 분석하므로써 동일 단어라도 PoS 별, 문장형태별로 다른 운률 생성 규칙을 적용해야만 합성음의 자연성을 확보할 수 있다는 사실을 보였다.

성능 평가 결과 약 52 %가 새로운 운률 규칙이 적용된 합성음이 보다 자연스럽다고 선택하였으며 약 27%가 차이가 없다고 판정하였다

이상에서 살펴본 바와 같이 자연어 처리기의 구문 분석 결과는 합성음의 자연성 개선에 무척 유용하게 이용될 수 있음을 알았다. 단 구문 분석 결과를 합성음의 자연성 개선에 효과적으로 사용하기 위해선 현재 결과 보다는 보다 일반화된 결과, 즉 운률 생성 규칙을 보다 다양한 음성 DB 에서 추출하는 것이 선결되어야 할 것이다.

참고 문헌

1. G. Borden and K. Harris, Speech science primer, Williams & Wilkins, 1980.

2. Wolfgang Hess, Pitch detection of speech signals. Springer-Verlag, 1983.

3. S. Furui and M. Sondhi, Advances in speech signal processing, Dekker, 1992.

4. D. Childers and A. Krishnamurthy, "A critical review of electroglottography," CRC Crit. Rev. Biomed. Eng. 12, pp.131-161, 1985.

5. D. Childers and M. Hahn, "Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech," IEEE Trans. Acoust., Speech, and Signal Process. ASSP-37, pp.1771-1774, 1989.

6. L. Rabiner and R. Schaffer, Digital Processing of speech signals, Prentice-Hall, 1978.