다양한 언어 정보를 이용한 음소 단위 억양 및 VoiceXML 문서 생성*

이화진[0] 박종철
한국과학기술원 전산학 전공 및 첨단정보기술연구센터
{matolee, park}@nlp.kaist.ac.kr

Diphone-based Intonation and VoiceXML document Generation using Multi-dimensional Linguistic Information

Lee Hwa Jin[O] Jong C. Park
Department of Computer Science, KAIST

요 약

최근 음성 합성 과정에서 화자의 의도를 가장 많이 반영하는 언어 정보인 문맥 정보를 사용하려는 시도가 이루어지고 있으나 문맥 정보를 적은 비중으로 사용하기 때문에 자연성 향상에 큰 도움을 주지 못하고 있다. 본 연구에서는 구문 정보, 의미 정보를 억양 생성 과정에 이용함과 동시에 문맥 정보와 음성 정보와의 관계를 음성 데이터를 바탕으로 분석하여 다양한 문맥 정보를 음성 합성 과정에 반영하는 방법을 제안한다. 또한 한국어에서 나타나는 다양한 억양 곡선 유형을 형태소를 이용하여 보다 효율적으로 처리할 수 있는 방법을 제안하여 자연스러운 억양 생성 시스템을 구현하고 시스템의 결과를 음소 단위 억양 생성기와 VoiceXML을 이용하여 적용시켜보고 결과를 논의한다.

1. 서론

키보드와 마우스만으로 컴퓨터에 일방적으로 정보를 제공하던 종래의 방식과는 달리 음성이 새로운 인간-기계 상호 작용의 도구로서 가지는 장점이 부각되면서 음성 언어 시스템에 관한 연구가 활발하게 진행되고 있다. 이 중 음성 합성 분야에서는 정확성과 자연성의 두 가지 기준을 충족시키기 위한 연구가 많이 진행되고 있는데 자연성에 있어서는 음의 길이, 휴지, 음의 높낮이 등이 부자연스럽게 생성되는 경우를 볼 수 있다. 인간은 같은 문장에 대해서도 문맥적 상황에 따라 전혀 다른 억양으로 발화할 수 있다. 예 (1)과 (2)에서 질문 문장에 따라 '빵을' 과 '영희가' 가 각각 응답 문장에서 중요한 정보이므로 이들을 강조하여 발화하게 된다.

(1) Q: 영희가 무엇을 먹었니?
A: 영희가 빵을 먹었어.

(2) Q: 누가 빵을 먹었니?
A: 영희가 빵을 먹었어.

이들 예는 질문-응답 형식의 간단한 예들이지만 낭독체의 복잡한 문장에 대해서는 문맥 정보를 이용하지 않고는 상황에 맞는 자연스러운 억양 정보를 생성해 낼 수 없다.

문맥 정보 외에도 구문 정보, 의미 정보 등이 음의 길이, 휴지, 음의 높낮이와 같은 억양 구성 요소의 변화에 영향을 준다. 본 연구에서는 억양 생성 과정에서 각 억양 요소에 영향을 주는 정보를 음성 데이터 분석을 통하여 분석하고 그 결과를 바탕으로 자연스러운 억양을 생성하는 방법을 보인다. 또한 이를 실제 음성 합성기에 적용시키는데 여러 가지 종류의 음성 합성기가 있으나 본 연구에서는 정교한 억양 요소 제어를 위하여 음소 단

* 본 연구는 첨단정보기술 연구센터를 통하여 한국과학재단의 지원을 받아 수행되었습니다.

위 연속 음성 합성기 (Diphone-based concatenative synthesizer)에 생성된 결과를 적용해 보고, 최근 음성을 이용하는 응용 시스템에 광범위하게 사용되고 있는 VoiceXML (Voice eXtensible Markup Language)을 이용한 억양 구성 요소 제어 가능성을 살펴본 후 음소 단위 억양 생성과 비교하여 장·단점을 분석한다.

2절에서는 억양 생성에 관한 관련연구를 살펴보고, 3절에서는 실제 음성 데이터를 바탕으로 억양 요소들과 여러 가지 언어 정보와의 관계를 분석한 후 억양 정보 생성 방법을 제안하고 4절에서는 분석 결과를 이용한 억양 생성 시스템의 구현을 보인다. 5절에서는 생성된 결과를 음소 단위 억양 생성기에 적용한 결과를 보이고 앞서 설명한 VoiceXML을 이용하여 억양 구성 요소를 제어할 수 있는 방법에 대하여 논의한 후 음소 단위 억양 생성과 비교하여 장·단점을 분석하고 6절에서 결론 및 향후 계획을 제시한다.

2. 관련연구

[1]은 담화 문맥에 의존하여 낭독체 담화문의 억양 생성 방법을 제안한다. 이 연구에서 주목할 점은 대조 강세 (contrastive stress)를 고려함으로써 보다 자연스러운 억양을 재현하고자 한 점이다.[1] 이 연구에서 억양 생성에 사용되는 문맥 정보인 Theme/Rheme은 previous-mentioned method에 의해 추출된다.[2] 이 방법에 따르면 명사, 동사 등의 내용어에 강세가 할당되고 기능어에는 할당되지 않는다. 또한 국지적 담화 분절 내에서 이미 나타난 단어에 대해서는 강세가 할당되지 않는다. 이러한 방법으로 강세를 생성하면 고강세를 과다하게 생성하여 청자가 화자의 의도를 오해할 가능성이 있고 [3] 대조 강세는 고려할 수 없게 되므로 이 연구에서는 이러한 문제점을 보완할 수 있는 알고리즘을 제안한다. 이 연구에서는 Theme/Rheme에 각각 강세 'L+H*' 와 'H*' 를 할당하는 것을 기본으로 하는데 영어에서도 상황에 따라 다양한 강세 패턴이 나타날 수 있으므로 상황에 맞는 다양한 억양 생성에 한계가 있다. 또 기능어에는 강세를 할당하지 않는다는 단점이 있다.

[4]에서는 유성음 수에 따라 억양 패턴을 생성하는 방법을 제안하는데 가능한 억양 패턴만을 생성하기 위해서는 효과적이지만 동일한 유성음 수의 어절이라도 문맥에 따른 억양의 상대적 변화를 다루기에는 어려움이 있다.

[5]에서는 합성음의 성능을 향상시키기 위해 자연 언어 생성 시스템으로부터 제공되는 언어 정보를 사용하는 SOLE Concept-to-Speech 시스템을 제안한다. 수작업으로 기본 주파수 곡선이 표기된 텍스트로부터 음절 단위로 언어 정보를 추출하고 이를 결정 트리 학습에 이용하여 결정 트리를 이용하여 강세를 할당한다. 문맥 정보는 명사구와 수사구조에 대한 것으로 수사구조로는 대조, 나열, 유사, 양보 등을 다루고 있고 명사구에 대해서는 통사적 유형으로 한정 명사, bare-singular[3], 명사 수식어로 나누고 의미적 유형으로 고유 명사와 상위어로 나누어 정보를 추출한다. 이러한 정보를 SGML (Standard Generalized Markup Language) 템플릿 형태로 구조화하여 억양 생성에 이용한다. 언어 정보로부터 SGML로의 자동 변환이 이루어진다면 보다 효율적으로 억양을 생성할 수 있을 것이다. 본 연구에서는 언어 정보를 XML 형식으로 처리하지는 않지만 XML을 이용하면 보다 체계적으로 언어 정보를 활용할 수 있을 것으로 보인다.

[6]에서는 한국어에서 기능어에도 억양 정보가 포함된다는 점에 착안하여 문맥 정보를 이용하여 한국어의 기능어인 조사에 강세를 할당하는 방법을 제안하였다. 이 연구는 [1]과 마찬가지로 영어의 Theme/Rheme에 대하여 각각 할당되는 'L+H*' 와 'H*' 를 한국어에 그대로 할당하여 한국어에서 나타나는 특징적인 억양 패턴을 고려하지 못하는 한계를 가지고 있다.

3. 억양 구성 요소의 생성

본 연구에서는 자료상에서 나타나는 억양 구성 요소들과 언어 정보들과의 관계를 분석하고 이를 바탕으로 억양 구성 요소 제어를 위한 규칙을 도출한다. 본 연구에서 분석에 사용한 데이터는 초등학교 말하기·듣기 교과서와 읽기 교과서의 내용을 녹음한 음성 데이터이다.[4] 음성 데이터 분석은 억양 분석 도구인 SFS (Speech Filing System)를 이용하여 수작업으로 수행하였다.

---

1) 대조 강세는 하나의 담화 개체가 다른 개체들과 구분될 수 있는 경우에 그 개체에 강세를 두어 발화하는 현상이다.

2) Theme이란 화자와 청자가 함께 이야기하기로 동의한 부분을 말하고 Rheme이란 새로운 부분을 말한다.[2]

3) 관사가 붙지 않는 단수 명사

4) http://mi.edunet4u.net/mullib

3.1 음의 길이 정보 생성

본 연구에서는 각 음소의 고유 길이를 기준으로 음의 길이 변화를 관찰하였다.[5] 표 1은 주요 모음의 고유 길이를 나타낸다. 음의 길이의 변화 중에서도 장음화에 초점을 두는데 실제로 음의 장음화가 일어나는 부분만 잘 조절하여도 훨씬 더 자연스럽게 들림을 알 수 있다.

모음	ㅏ	ㅣ	ㅡ	ㅗ	ㅓ	ㅜ
고유길이	107.2ms	84.9ms	81.1ms	114.6ms	106.2ms	125.3ms

표 1 주요 모음의 고유 길이

3.1.1 구문 정보와 음의 길이

데이터 분석 결과 음의 길이에 영향을 주는 정보 중에서 가장 많은 영향을 미치는 정보는 구문 정보이다.

(3) 강아지와 토끼가(270ms) 커다란 나무 밑에서(176ms) 널뛰기를 하고있습니다.

예 (3)에서 음소 'ㅏ' 와 음소 'ㅓ' 의 고유 길이가 각각 107.2ms와 106.2ms인데 비하여 밑줄 친 음절에서의 해당 음소의 길이가 각각 약 163ms와 70ms 만큼 길게 발음되었다. 데이터 분석 결과, 부사구가 삽입되는 경우에 부사구의 앞과 뒤에서 장음화 현상이 일어났다.

(4) 돼지가(250ms) 널을 힘차게 구르자(175ms) 토끼는 붕 날아올라(166ms) 나뭇가지에 걸리고 말았습니다.

예 (4)에서는 주어와 목적어의 경계, 절과 절의 경계에서 장음화 현상이 일어난다. 표 2는 문장 구조에 따른 장음화 현상을 규칙화 하여 표로 정리한 것이다.

3.1.2 의미 강조 및 단어의 특성과 음의 길이

표 2에 기술한 규칙 외에도 장음화 현상이 나타난 경우가 있었으나 이들의 경우는 규칙으로는 일반화하기 어려운 점이 있다.

문장 구조	장음화 규칙
주어+부사구+목적어+서술어	주어 마지막 음소 부사구 마지막 음소
부사구+주어+목적어+동사	부사구 마지막 음소 주어 마지막 음소
주어1+목적어+동사1+주어2+동사2+동사3	주어 마지막 음소 동사1 마지막 음소 동사2 마지막 음소
부사(주어역할)+목적어+서술어	부사 마지막 음소
부사+주어+목적어+동사1+동사2	부사 마지막 음소 주어 마지막 음소 동사1 마지막 음소
주어1+동사1+동사2	동사1 마지막 음소
주어+부사+서술어	부사 마지막 음소

표 2 문장 구조와 장음화 규칙

(5) 따뜻(171ms)한 봄날입니다.

예 5에서 '따뜻한' 의 음소 'ㅡ' 에 장음화가 일어나는 것은 동화 도메인에서 특징적으로 나타나는 특정한 현상으로 보인다. 즉, '따뜻하다' 는 느낌과 봄날의 '나른함' 을 강조하기 위하여 길게 발음하는 것이다.
다음은 단어의 특성에 의한 장음화의 예를 보인다.

(6) (a) 돼지가 널(141ms)을 힘차게 구르자 토끼는 붕 날아올라 나뭇가지에 걸리고 말았습니다.
(b) 강아지와 토끼가 커다란 나무 밑에서 널(93ms) 뛰기를 하고있습니다.

'널' 이라는 동일한 단어에 대하여 단독으로 사용되는 경우에 합성 명사 내에서 사용되는 경우보다 길게 발음되는 경향이 나타난다. 이러한 현상은 동화 도메인과 같이 발화시에 발화자의 감정이 이입되는 도메인에서 나타나는 특수한 현상이다. 또한 단어의 특성과 문장 내 문맥, 발화자의 발화 습관에 따라 특징적으로 나타나는 현상이기도 하므로 표 2에 보인 규칙과 같은 형태로 일반화 하기는 어렵다. 모든 도메인에 맞는 보편적인 억양 생성 규칙을 만들기는 거의 불가능하므로 각 도메인마다 나타나는 이러한 예외적인 현상들을 처리할 수 있다면 보다 자연스러운 음성 합성 결과를 기대할 수 있다.

3.2 음의 높낮이 정보 생성

억양의 자연스러움에 가장 많은 영향을 미치는 억양 구성 요소는 음의 높낮이이다. 본 연구에서는 한국어의 문

5) 고유 길이란 해당 음소가 문맥 없이 단독으로 발음될 때의 음의 길이를 가리키지만 수집된 음성 데이터의 발화자가 직접 고유 길이를 녹음할 수 없는 제약으로 인하여 해당 데이터 내에서 각 음소별 평균 길이를 구하여 이를 고유 길이로 대신 사용하였다.

맥 정보, 구문 정보, 의미 정보와 음의 높낮이 정보와의 관계를 분석하여 그 결과를 자연스러운 억양 생성에 이용한다.

3.2.1 형태소별 음의 높낮이

형태소는 뜻을 갖는 최소의 단위이다 [4]. 영어에서는 각 단어마다 고유의 강세와 높낮이 유형을 가지는데 데이터 분석 결과 한국어에서는 형태소 단위로 고유의 높낮이 패턴이 나타났다. 그림 1은 본 동화 데이터에서 나타나는 주요 모음의 음높이 유형이다.

높낮이 유형	높낮이 곡선
상승	
하강	
상승-하강	
하강-상승	
유지	

그림 1 음소별 음높이 유형

그림 1의 분류는 음소별 분류에 해당하나 형태소는 곧 음소의 연속이므로 형태소 단위로도 음높이가 유형화될 수 있다. 특히 명사의 경우에 그 특징이 두드러진다.

(7) (a) 강아지와 토(352Hz-249Hz)끼(249Hz-149Hz)가...
(b) 돼지와 토(357Hz-288Hz)끼(288Hz-195Hz)가 ...
(c) 돼지가 널을 힘차게 구르자 토(340Hz-235Hz)끼 (235Hz-195Hz)는 ...
(d) 토(336Hz-226Hz)끼(226Hz-157Hz)는 ...

예 (7)의 각 문장에서 '토끼' 의 음의 높낮이 유형은 하강형에 해당한다.[6] 그러나 각 문장에서의 '토끼' 의 높낮이 유형은 동일하나 시작점의 음높이와 시작 음소와 마지막 음소의 음높이 차이는 각기 다름을 알 수 있다. 따라서 대략적인 높낮이 유형만으로는 상황에 따라 변하는 정확한 음의 높낮이를 생성하기는 어렵다.

3.2.2 문맥 정보와 음의 높낮이

음높이를 결정하는 요소들 중 발화자의 의도와 가장 밀접한 관련이 있는 요소는 문맥 정보이다. 다양한 문맥적 상황에서 나타나는 억양을 자연스럽게 생성해 내기 위해서는 억양 생성 시에 문맥 정보를 이용해야 한다. 문맥 정보에는 여러 가지가 있을 수 있는데 억양은 정보의 새로움에 영향을 많이 받는다. 관련연구에서 논의한 것과 같이 화자는 발화시에 청자에게 새로운 정보에 대해서는 높은 음높이를 할당하여 발화함으로써 전달하려는 정보를 강조하게 되어 정보의 중요도가 높음을 암시한다. 이와 같은 종류의 문맥 정보는 화제-논평 구조로 설명할 수 있다. 화제란 이미 논의한 테마 (theme), 전제 (presupposition)의 의미를 포함하는 것으로 화자의 입장에서 청자의 기억 속에 이미 저장되어 있다고 예상되거나 저장되어 있는 지식으로부터 추론 가능한 정보를 말한다. 추론 가능한 정보가 화제가 되어야 하는 것은 이미 문맥상에 나타났던 정보라도 다른 표현으로 재등장할 수 있기 때문이다.

(8) (a) 옛날 중국 어느 마을에 경치가 좋은 산(372Hz) 이 있었어요.
(b) 그 산(310Hz)에는 소를 치는 소년이 ...

예 (8)에서 '산' 이라는 단어가 (a)에서는 새로운 정보이고 (b)에서는 주어진 정보가 되어 음높이 값의 차이를 보인다. 이러한 현상은 다음 예에서도 찾아볼 수 있다.

(9) (a) 그 산에는 소를 치는 소(373Hz)년(308Hz)이 ...
(b) 소(324Hz)년(281Hz)은 어느 날 소를 몰고 가다 ...

이와 같이 문맥적 상황에 따라 동일한 단어에 대하여 시작 음높이가 달라짐을 알 수 있다. 시작 음높이 외에 문맥 정보에 따라 형태소의 발화 시작점으로부터 발화 끝점 사이의 음높이 차이도 달라진다.

(10) (a) 강아지와 토(352Hz-249Hz)끼(249Hz-149Hz)가...
(b) 다음에는 돼지와 토(357Hz-288Hz)끼(288Hz-195Hz)가...

'토끼' 는 (10)의 (a)와 (b)에서 각각 논평과 화제에 해당하는데 음높이 차이는 각각 203Hz와 152Hz로 이미 주어진 정보에 대하여 음 높이의 하강 폭이 줄어듦을 볼 수 있다. 이는 화자가 담화를 이끌어 나가면서 해당 단어에 대하여 정보의 중요도와 관심도를 낮게 할당하기 때문이라고 생각할 수 있다. 이러한 규칙은 문맥 윈도우 (context window)와 관련지어 생각해야 하는데 문맥 윈도우란 단락 구분에 의한 것이기보다는 의미상의 구분을 말한다.[7] 현재 문맥 윈도우에서 이미 주어진 정보이며 시간이 지날수록 중요도와 관심도가 떨어지는 표현에 대

6) 괄호 안의 값은 음성 정보 분석 도구인 Praat으로 측정한 기본 주파수이다.

해서는 시작 음높이가 낮아지고 그와 동시에 발화 시작점과 발화 끝점의 음높이 차이도 줄어든다. 그러나 의미상의 전환에 의한 새로운 문맥 윈도우가 시작되면 이전 문맥 윈도우에서는 이미 주어진 정보로 중요도 및 관심도가 낮은 표현에 대해서도 다시 발화 시작점과 발화 끝점의 음높이 차이가 커진다.

(11) (a) 돼지가 널을 힘차게 구르자 토(340Hz-235Hz)끼(235Hz-195Hz)는 붕 날아올라 나뭇가지에 걸리고 말았습니다.
(b) 마침 지나가던 기린이 이 모습을 보고 뛰어왔습니다.
(c) 토(336Hz-226Hz)끼(226Hz-157Hz)는 기린의 목을 타고 땅으로 내려왔습니다.

예 (11)에서 (a)까지 이전 문맥 윈도우가 유지되고 (b)에서 '마침' 이라는 단어와 새로운 캐릭터인 '기린' 이 등장함으로써 새로운 문맥 윈도우가 시작됨을 알리고 (c)의 '토끼' 는 시작점과 끝점의 차이가 179Hz로 (a)의 145Hz보다 다시 커지는 것을 볼 수 있다.

3.2.3 구문 정보와 음의 높낮이

앞서 설명한 문맥 정보 외에 구문 정도 또한 음의 높낮이를 결정하는데 영향을 준다. 어휘의 문장 내 위치가 음의 높낮이에 영향을 줄 수 있다.

(12) (a) 다음에는 돼지와 토(357Hz-288Hz)끼(288Hz-195Hz)가 널뛰기를 합니다.
(b) 돼지가 널을 힘차게 구르자 토(341Hz-235Hz)끼(235Hz-195Hz)는 붕 날아올라 나뭇가지에 걸리고 말았습니다.
(c) 토(336Hz-226Hz)끼(226Hz-157Hz)는 기린의 목을 타고 땅으로 내려왔습니다.

'토끼' 라는 단어는 예 (12)의 (a),(b),(c)에서 각각 세 번째, 다섯 번째, 첫 번째 성분에 해당하는데 시작 음높이와 끝 음높이의 차이는 (c)>(a)>(b)의 순임을 볼 수 있다. 한국어에서는 일반적으로 문두에서 문미로 갈수록 절대적인 음높이가 낮아지는데 이 때 어휘 내에서의 변화폭 역시 줄어드는 경향을 보인다. 따라서 정보의 중요도와 음높이는 밀접한 관계가 있음을 알 수 있다.

조사의 경우, 항상 음의 높낮이 유형이 동일하지 않고 상황에 따라 달라지는데 이는 구문정보와 관계가 있다.

(13) (a) 강아지와 토끼가(202Hz-296Hz-155Hz) <P> 커다란 나무 밑에서 널뛰기를 하고있습니다.
(b) 돼지가(222Hz-313Hz-50Hz) <P> 그 옆에서 구경을 합니다.

예 (13)에서 조사 '가' 의 음높이 유형은 상승-하강 형태이다. 두 문장 모두 <주어>+<부사구>+<목적어>+<서술어>의 형태로 '가' 는 주격조사에 해당한다. 한국어의 일반적인 문장 구조인 <주어>+<목적어>+<서술어>의 형태에 <부사구>가 삽입되면 주어와 부사구 경계에 휴지가 삽입되고 휴지 바로 앞 음소는 상승-하강 형태를 보이는 경우가 많다. 이 경우 주어의 마지막 음소는 주격 조사인 경우가 대부분이므로 조사의 음높이 유형과 구문정보와의 관계가 밀접함을 알 수 있다. 이와 같이 한국어에서는 체언과 같은 내용어 뿐만 아니라 조사와 같은 기능어에도 음의 높낮이 정보가 특징적으로 나타난다.

3.3 휴지 정보 생성

실제 화자는 호흡 조절을 하기 위해 혹은 정보 전달 효과를 높이기 위해 적절한 위치에 휴지를 삽입한다.[8] 휴지 현상이 일어나는 위치는 구문 정보에 영향을 받는데 그중에서도 문장의 구조에 많은 영향을 받는다.

(14) (a) 강아지와 토끼가 <P> 커다란 나무 밑에서 <P> 널뛰기를 하고있습니다.
(b) 돼지가 <P> 그 옆에서 <P> 구경을 합니다.

예 (14)에서는 앞 절에서 설명한 것과 같이 부사구가 삽입되었을 때 주어와 부사구 경계에서 휴지가 나타나고 부사구의 끝 지점에서 다시 한번 휴지가 나타난다. 부사구가 문장 맨 앞으로 나오는 경우에도 부사구 끝 지점에서 휴지가 나타난다.

(15) (a) 다음에는 <P> 돼지와 <P> 토끼가 <P> 널뛰기를 합니다.

예 (15)에서는 주어가 병렬 구조를 이루는데 이때 두 어절 사이에 휴지가 일어난다.[9] 종속문의 예를 보면,

7) 동화 도메인에서 문맥 윈도우란 하나의 사건이 유지되는 담화 분절을 의미한다.

8) 기존 TTS 시스템을 통하여 생성된 발음이 부자연스럽게 들리는 이유 중 하나가 부적절한 위치에 휴지를 삽입하기 때문으로 보인다.

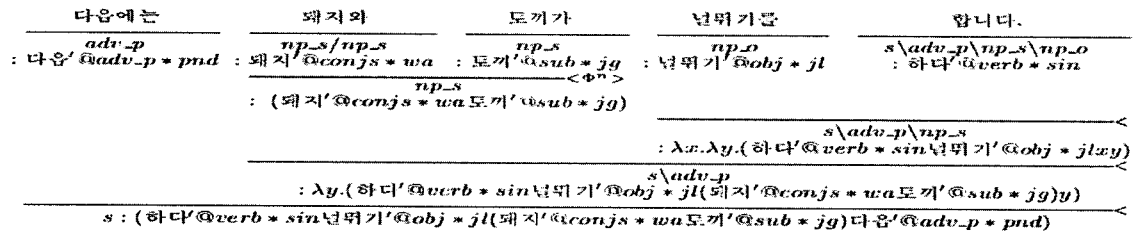

그림 3 결합범주문법을 통한 문장 분석

(16) (a) 돼지가 <P> 널을 힘차게 구르자 <P> 토끼는 <P> 붕 날아올라 <P> 나뭇가지에 걸리고 말았습니다.

예 (16)은 주어가 두개이고 서술어가 두개 이상인 종속문의 경우로 종속절과 주절의 주어 다음에 휴지가 나타나고 절과 절의 경계에서 휴지 현상이 나타남을 알 수 있다.

4. 억양 정보 생성 시스템의 구현

본 절에서는 앞서 설명한 억양 구성 요소와 언어 정보와의 관계를 바탕으로 억양 정보 생성 시스템의 구현을 보인다. 시스템의 개요는 그림 2와 같다. 먼저 자연 언어 문장을 입력 받아 결합범주문법을 통하여 분석한다. 그림 3은 결합범주문법을 통한 문장 분석 과정이다. 분석된 결과를 기반으로 문맥 정보를 추출한다. 본 논문에서는 화제-논평 정보를 이용하는데 이를 추출하기 위해 previous-mentioned method와 word class 개념을 이용한 추론을 사용한다.

(17) (a) 애벌레 푸는 정원에 사는 모든 곤충들의 친구랍니다.
(b) 정원에 사는 메뚜기는 푸의 친구예요.

예 (17)에서 (b)에서 '메뚜기' 는 (a)에 나타난 '모든 곤충들' 에 속하므로 새로운 표현이지만 추론 가능한 화제에 속한다. 따라서 유사한 표현들을 하나의 상위어 혹은 대표어로 묶어서 분류하는 방법을 사용하면 이와 같은 문제를 해결할 수 있다. 이러한 방법으로 문맥 정보를 추출하면 다음과 같다.

9) <주어>+<목적어>+<서술어> 형태에서 보통 주어 뒤에 휴지가 온다.

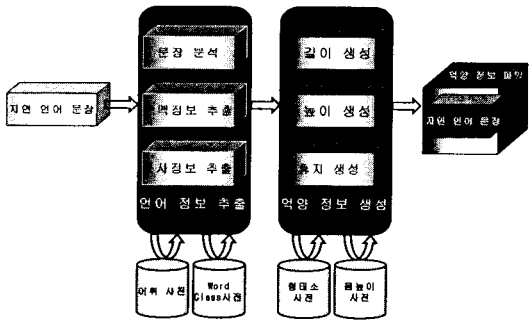

그림 2 억양 생성 시스템의 구조

- [다음에는]논평 [돼지와 토끼가 널뛰기를 합니다.]화제

다음으로 구문 정보를 이용하여 표 2의 규칙에 따라 음길이를 생성하고 휴지 정보를 생성한다.[10] 음길이 정보와 휴지 정보를 생성한 결과는 다음과 같다.

- 다음에는(283ms)<P:390ms>돼지와<P:58ms>토끼가(137ms)<P:396ms> 널뛰기를 합니다.

음의 높낮이 정보는 형태소별로 제어되는데 이때 각 형태소는 PHO 파일에서 사용되는 SAMPA 표기법을 따라 표시된다.[11] 어절별 형태소 사전은 다음과 같이 구성된다.

- tag('돼지@sub*jg,[['t],'wE','tsW','i'],['k','a']])

마지막으로 문맥 정보와 구문 정보를 이용하여 각 형태소별로 음높이 정보를 생성한다.

10) 장음화가 일어나지 않는 음소에 대해서는 코퍼스 내에서 계산된 음소별 평균 길이를 적용하였다.
11) http://asadal.cs.pusan.ac.kr/hangeul/hndb/hn13smp.html

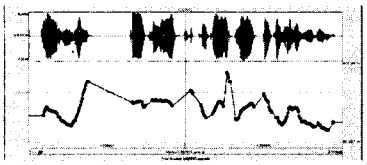

그림 6 원음의 분석 결과

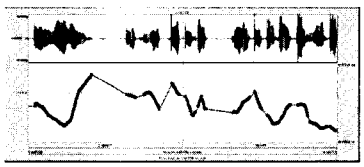

그림 7 합성음의 분석 결과

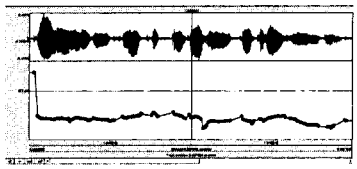

그림 8 타 TTS의 합성 결과

5. 음소 단위 억양 생성 결과와 VoiceXML과의 비교

본 절에서는 앞서 설명한 억양 정보 생성 시스템의 결과를 음소 단위 연속 음성 합성기에 적용하고 VoiceXML을 이용한 억양 제어 방법을 제시한 후 음소 단위의 음성 합성과 VoiceXML을 이용한 음성 합성을 비교하여 장·단점을 분석한다.

5.1 음소 단위 억양 생성기에 적용한 결과

본 연구에서는 생성된 억양 정보를 음소 단위 연속 음성 합성기인 MBROLA에 적용해 보고 이를 통하여 억양 정보 생성 시스템의 결과를 평가한다.[12] MBROLA를 이용하여 음성을 합성하기 위해서는 MBROLA에 입력으로 PHO 파일을 제공해야 한다. 그림 4는 문장 "다음에는 돼지와 토끼가 널뛰기를 합니다." 에 대하여 생성된 억양 정보를 포함하는 PHO 파일이다. 각 행의 첫 번째 열은 음소를 나타낸다. '_' 는 휴지를 나타낸다. 두 번째 열은 각 음소의 길이를 나타내고 세 번째 열부터는 음의 높낮이를 표시하는데 세 번째 열은 해당 음소의 전체 길이 내에서의 비율을 나타내고 네 번째 열은 해당 구간 내에서의 음높이를 Hz 단위로 나타낸다. 다섯 번째 열부터는 다시 음높이 정보가 반복된다. 위와 같이 생성된 PHO 파일을 음소 단위 연속 음성 합성기인 MBROLA에 적용한 결과를 Praat을 이용하여 분석하였다. 테스트에는 초등학교 교과서에 수록된 동화 <널뛰기> 중 여섯 문장이 사용되었다. 그림 6과 그림 7은 각각 문장 "다음에는 돼지와 토끼가 널뛰기를 합니다." 에 대한 원음과 합성음의 분석 결과를 나타낸다. 분석 결과는 두 부분으로 나누어지는데 윗 부분은 음의 세기이고 아래 부분은 억양 곡선을 나타낸다. 본 연구에서는 PHO 파일에 음의 세기를 제어할 수 있는 부분이 없으므로 음의 세기는 고려하지 않는다. 억양 곡선의 형태를 보면 그림 6과 그림 7이 거의 비슷한 형태를 보이는데 표시된 부분은 원 문장에서 호흡 부분으로 본 연구에서 미세한 호흡까지 고려하지 않았기 때문에 합성음에서는 이 부분이 나타나지 않았다. 음의 높낮이 면에서 합성음의 음의 높이가 전체적으로 약 30Hz 정도 낮음을 알 수 있는데 이것은 분석된 음성 코퍼스는 여성 발화자에 의해 녹음된 것이고 MBROLA의 음소 데이터베이스는 남성 발화자에 의해 녹음된 것이기 때문에 이를 보정하기 위하여 생성 시에 30Hz 낮게 생성하였기 때문이다. 음의 길이 면에서 나타나는 차이는 본 연구에서는 장음화가 일어나는 음소 외에 다름 음소에 대해서 코퍼스 내에서의 해당 음소의 평균 길이를 사용하였기 때문으로 분석된다. 그림 8은 문맥 정보를 고려하지 않은 TTS 시스템의 합성음 분석 결과이다. 본 연구의 결과는 억양 곡선 상으로는 원음과 유사하나 합성음을 실제로 들어보면 아직 부자연스러운 점이 많은데 그 원인에는 여러 가지가 있을 수 있으나 우선 여성 발화자의 데이터를 기반으로 생성된 억양 정보를 남성 발화자의 목소리로 발음을 하여 주파수 대역의 차이로 인해 어색하게 들린다고 볼 수 있다. 또한 음의 높낮이 정보를 코퍼스 내에서의 각 음소별 평균값을 기준으로 변화시킴으로써 유발되는 음소 간 불연속성을 어색함의 원인으로 들 수 있다. 이러한 문제점을 해결하기 위해서 하나의 음소가 주변 음소에 의해 받는 영향을 고려하여 음소 간의 음높이 보정이 필요하다.

5.2 VoiceXML을 이용한 억양 구성 요소 제어

앞서 보인 음소 단위 음성 합성기 이외에 최근 음성 응용 프로그램의 제작 도구로 많이 사용되고 있는 VoiceXML의 speech markup을 통하여 억양 구성 요소를

12) MBROLA는 연구 목적으로 만들어진 음성 합성기로 현재 25개 국어에 대하여 남성 및 여성 음성으로 녹음된 음소 데이터베이스를 제공하는데 한국어에 대해서는 남성 음성으로 녹음된 음소 데이터베이스 hanmal을 제공한다. hanmal 음소 데이터베이스는 http://asadal.cs.pusan.ac.kr/hangeul/hndb에서 제공된다.

제어할 수 있다. 휴지, 음의 세기, 음의 높낮이, 음의 길이 외에도 음높이 곡선 유형, 발화 속도, 출력 크기 등을 조절할 수 있고 발화자의 성, 나이 등의 억양 이외의 정보를 억양 생성에 이용할 수 있다.[13] 그러나 현재 한국어에 대하여 VoiceXML을 사용하여 음성 인식 및 음성 합성 응용 프로그램을 개발할 수 있는 무료 소프트웨어는 없는 실정이다.[14] 이에 따라 영어를 지원하는 IBM의 VoiceServer SDK를 이용하여 한국어 억양 생성에 관한 실험을 선행 실험을 수행하였다. 그림 9는 VoiceXML Speech Markup을 이용하여 억양 정보를 제어한 예의 일부이다. 그림 9에서 <ibmlexicon>~</ibmlexicon> 부분은 각 음절에 대하여 IPA 표기법에 따라 한국어에 맞게 발음을 재 정의한 부분이다. 이와 같이 VoiceXML을 사용하면 음소 데이터베이스를 따로 구축하지 않아도 다양한 방법으로 간단히 발음을 조정할 수 있다.[15] 그러나 합성된 결과를 들어보면 한국어 음성 합성기가 아니기 때문에 아직도 외국인이 발음하는 것과 같이 부자연스러운 느낌을 준다. VoiceXML을 이용하면 음소 레벨에서의 정교한 억양 제어는 불가능하지만 VoiceXML 2.0에서 제공하는 다양한 속성들을 이용하면 보다 자연스러운 억양을 생성할 수 있을 것으로 보인다.[16]

```
입력: 다음에는 돼지와 토끼가 널뛰기를 합니다.
<ibmlexicon>
<word spelling="da" pronunciation="d&#593;"/>
<word spelling="um" pronunciation="&#650;m"/>
<word spelling="e" pronunciation="&#603;"/>
<word spelling="neun" pronunciation="nun"/> . . .
</ibmlexicon>
<form>
<block>
    <prompt>
          <pros duration="219" pitch="208.33"> da </pros>
          <pros duration="206" pitch="187.5"> um </pros>
          <pros duration="158" pitch="134.5"> e </pros>
          <pros duration="473" pitch="258.75"> neun </pros>
          <break time="390"/>
```

그림 9 VoiceXML을 이용한 억양 제어의 예

6. 결론 및 향후 계획

본 논문에서는 보다 자연스러운 억양 생성을 위하여 구문 정보와 의미 정보 외에 억양 변화에 중요한 영향을 미치는 문맥 정보를 이용하였고 그 결과를 음소 단위 연속 음성 합성기에 적용하였다. 또한 음성 응용 프로그램 제작에 광범위하게 사용되고 있는 VoiceXML을 이용하여 억양을 제어할 수 있는 방법을 제안하였다. 본 연구에서는 문맥 정보 중에서 정보의 새로움에 주안점을 두고 음의 높낮이 정보를 생성하였으나 정보의 새로움 이외에도 억양에 영향을 주는 문맥 정보가 존재하므로 이를 분석하여 억양 생성에 사용하면 보다 자연스러운 억양을 생성할 수 있을 것으로 보인다. 한국어 지원 VoiceXML 개발 환경이 갖추어 지는 대로 VoiceXML에 대한 실험을 완결하여 본 논문에서 제시한 음소 단위 합성 시스템에 비하여 향상된 시스템을 구현할 예정이다.

7. 참고 문헌

[1] S. Prevost. 1995. Contextual Aspects of Prosody in Monologue Generation. IJCAI.

[2] M. Steedman. 2000. Information structure and the syntax phonology interface. Linguistic Inquiry.

[3] A. Monaghan. 1991. Intonation in Text-to-Speech Conversion System. Ph.D. thesis, University of Edinburgh.

[4] S. Kim and C. Kim. 1993. A study on the phonetic/prosodic rule based on the morphological analysis. The Magazine of the IEEK.

[5] J. Hitzeman et al. 1998. On the Use of Automatically Generated Discourse-level Information in a Concept-to-Speech Synthesis System. ICSLP.

[6] H. Lee and C. Park. 2000. Computational Generation of Context-based Intonation for Korean with Combinatory Categorial Grammar. ICCPOL.

13) voice markup의 속성 gender를 사용하여 여성 음성으로 합성하면 음소 데이터베이스가 남성 발화자의 음성이었기 때문에 생긴 부자연스러움을 해결할 수 있다.

14) 한국어를 지원하는 유료 소프트웨어는 몇 가지가 있으나 유료 소프트웨어를 이용할 경우 개발 환경에 의존적인 시스템이 개발될 수 있으므로 무료 소프트웨어 사용을 지향한다. IBM사의 WebSphere VoiceServer SDK가 이와 같은 기능을 제공하지만 한국어에 대해서는 아직 지원되지 않고 있다.

15) 같은 단어에 대해서 여러 가지로 녹음하여 사용할 수 있기 때문에 방언의 처리에도 VoiceXML의 사용이 가능할 것으로 보인다.

16) VoiceXML을 이용하면 미리 녹음된 사운드 파일을 재생시킬 수 있는데 한국인 발화자에 의하여 음절별로 녹음된 재생시키는 방법을 사용하면 음절 단위의 자연성은 보장할 수 있지만 파일을 읽는 시간 때문에 음절 간 불연속성이 나타나게 된다.