Stem-ML에 기반한 한국어 억양 생성

한영호(KT), 김형순(부산대)

<차 례>

<Abstract>

Korean Prosody Generation Based on Stem-ML

Young Ho Han, Hyung Soon Kim

In this paper, we present a method of generating intonation contour for Korean text-to-speech (TTS) system and a method of synthesizing emotional speech, both based on Soft template mark-up language (Stem-ML), a novel prosody generation model combining mark-up tags and pitch generation in one. The evaluation shows that the intonation contour generated by Stem-ML is better than that by our previous work. It is also found that Stem-ML is a useful tool for generating emotional speech, by controling limited number of tags. Large-size emotional speech database is crucial for more extensive evaluation.

* Keywords: Text-to-speech, Prosody, Stem-ML, Emotional speech synthesis.

1. 서 론

억양은 동일문장일지라도 그 문장이 어떤 맥락에서 이야기되는지에 따라, 또한 발화자가 무엇을 강조하고자 하는지에 따라 다르게 표현된다. 따라서 자연스러운 음성합성을 위해서는 동일문장일 경우라도 표현하고자 하는 의도, 강조점 등에 따라 적절한 억양패턴을 생성할 수 있는 억양 모델을 개발하여야 한다. 기존의 억양 생성 모델들은 이 문제에 대한 효과적인 해결책을 가지고 있지 못하다. 본 연구에서는 이러한 문제를 극복하기 위해 새로운 억양 생성 모델로 제시되고 있는 Soft TEMplate Mark-up Language (Stem-ML)[1] 방식을 도입하였다. 기존의 억양 모델과 비교하여 Stem_ML 방식의 가장 큰 장점은 마크업 태그(mark-up tags)와 억양생성을 하나의 시스템으로 통합함으로써, 예를 들어 강조나 대조로 인해 돋들림이 발생할 경우, 이를 모델링하기 위해 각 억양 패턴의 참조틀(template)에 각기 다른 강도값(strength tag)를 설정할 수 있다는 점이다. 이미 Stem-ML방식을 이용하여 영어와 중국어에 대한 억양 모델이 제작되었으며 그 결과 매우 자연스런 억양을 생성할 수 있었다[2][3]. 본 연구에서는 한국어 억양 생성에의 적용을 위해 기존의 Stem-ML 모델을 한국어에 맞게 수정하였다.

Stem-ML에 기반한 억양 생성 방식은 정서음성 합성을 위한 효과적인 도구로 활용될 수 있다. Stem-ML은 Speech Synthesis Mark-up Language (SSML)의 억양 모델 설계 방식을 따르고 있어서, 여러 단계의 언어 수준에 따라 각각 하위 태그(tag)를 이용하여 억양의 생성 방식을 조절할 수 있다. 뿐만 아니라 단순히 태그값을 변경시킴으로써 전체 억양의 패턴을 쉽게 수정할 수 있다. 이러한 Stem-ML의 유연성은 인간이 가진 정서적 특성을 모사할 때 큰 장점으로 작용할 것으로 판단하였다.

본 연구는 정서음성 합성을 위한 억양 생성에 관한 예비 연구의 성격을 띤다. 즉 Stem-ML 에서 제공하는 태그값들을 변경하여 각 정서음성의 억양을 생성함으로써 Stem-ML이 향후 정서음성 합성에 유용한 억양 생성 모델인지를 검증하는데 그 목적이 있다.

이상 본 연구의 목적은 크게 두 가지로 요약된다. 첫째, 한국어에 적합한 억양 생성 모델을 제작하기 위해 Stem-ML 방식을 한국어에 맞게 수정, 적용하여 그 가능성을 검증하고자 한다. 둘째 향후 정서 음성 합성을 위한 억양 생성 모델로서의 Stem-ML의 적합성을 타진하고자 한다.

본 논문의 구성은 다음과 같다. 서론에 이어 2장에서 Stem-ML의 전반적인 개요를 다룬다. 3장에서는 이러한 Stem-ML을 실제 한국어에 적용하기 위해 사용한 여러 가지 기법들을 살펴보고, 4장에서 이에 대한 실험 및 결과에 대해 기술한다. 5장에서는 Stem-ML을 이용한 정서음성 합성을 위한 예비적인 시도들에 대해 기술하고, 이에 대한 실험결과도 언급한다. 마지막으로 6장에서 결론을 맺는다.

2. Stem-ML의 개요

2.1. Stem-ML의 기본 개념

Stem-ML의 기본 개념은 실제 인간의 발화행위의 경험적 데이터를 기반으로 만들어 졌다[1]. 이러한 개념에는 선행계획(pre-planning), 억양구 간의 비상호작용, 피치 곡선의 연속성, 물리적 에너지와 발음의 정확성 간의 trade-off 등이 포함된다. 선행계획은 인간 발화의 인지적 특성을 반영하는 개념으로서, 사람은 말을 할 때 물리적인 노력을 최소화하는 방향으로 미리 음성발화의 계획을 세워둔다는 것이다. 하지만 이와 같은 노력의 최소화는 듣는 이로 하여금 자신의 발화를 이해하는데 장애가 되어서는 안 된다. 따라서 전체발화의 중요한 부분을 강하게 발화하고 나머지 부분은 될 수 있는 한 약하게 발화하는 방향으로 계획을 세우게 되는데, 강하게 발음해야 하는 음절이 나타나면 그 전의 약한 음들은 더욱 약하게 발음하며, 영향을 받은 약한 발음은 원래 자신이 가진 억양의 특성을 잃어버리게 된다. 즉 하나의 억양패턴의 모양과 강도는 그와 인접한 다른 억양패턴의 모양과 강도에 따라 달리 발화된다[1].

근육의 위치를 조절하는데 필요한 물리적인 에너지와 발음의 정확성을 위한 tone/accent 참조틀과의 대응에는 trade-off가 존재한다. 이는 인간 발화의 기본적인 목적, 즉 자신의 의사를 정확히 상대에게 전달하고자 하는 목적과 이를 실현하기 위한 근육의 운동량은 최소화하고자 하는 목적이 서로 상충되기 때문에 발생한다.

이와 함께 Stem-ML에서는 사람들이 하나의 억양구 안에서만 선행계획을 할 뿐이라고 가정하고 있다. 즉 하나의 억양구가 끝나고 다음 억양구가 시작될 때에는 앞의 억양구 패턴이 뒤의 억양구 패턴에 어떤 영향도 주지 않는다고 가정한다.

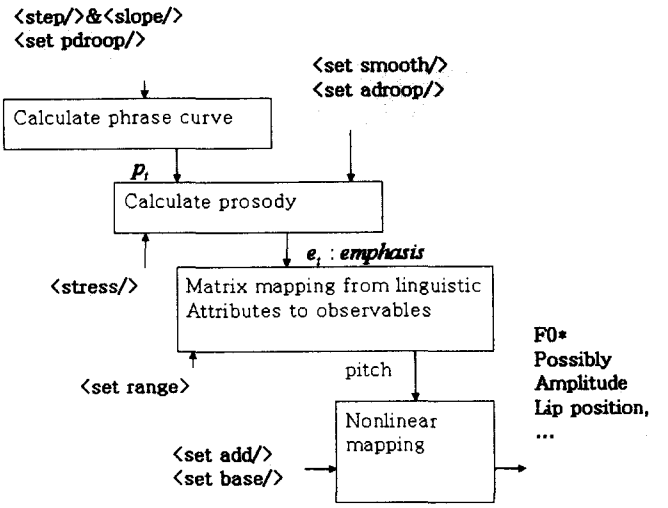
의식적으로 인간의 근육은 100ms 보다 빨리 반응할 수 없다. 100ms는 일반적으로 하나의 음절보다 긴 구간이기 때문에 음절들 사이에도 높은 상호작용이 존재하며 Stem-ML는 이를 근거로 억양 곡선은 매우 연속적이며 부드럽게 유지된다고 가정하고 있다.

2.2. Stem-ML의 수학적 기반

Stem-ML의 단위 시간은 통상 10ms이고 억양곡선은 각 단위시간에 피치 정보를 포함하고 있는 선형방정식을 설정하고 이를 풀어나감으로써 구할 수 있다. 이러한 선형방정식은 몇 개의 그룹으로 나눌 수 있는데, 첫번째 그룹은 피치곡선의 부드러움의 정도와 연속성을 표현하는 방정식들이다. 여기에 피치곡선에 대한 제한을 주기 위한 태그들로부터 파생되는 또 다른 방정식이 있다. 만일 이 두 그룹이 포함하고 있는 방정식들을 모두 만족하는 피치 곡선이 존재하지 않는다고 해

도, Stem-ML은 최소평균자승법(least-mean-square solution)을 이용해 주어진 방정식들에 가장 부합하는 피치곡선을 도출해 낸다. 이렇게 정확한 제약조건이 없이도 주위의 환경의 제약조건을 통해 유연하게 피치곡선을 그려내는 것이 Stem-ML의 가장 큰 특징이라고 할 수 있다.

Stem-ML에서 피치곡선을 도출해 내는 알고리즘은 <그림 1>에서 볼 수 있는 바와 같이 크게 네 단계로 이루어져 있다. 보다 자세한 사항은 참고문헌 [1]에서 찾을 수 있다.



<그림 1> Stem-ML의 블록도 [1]

3. Stem-ML 기반의 한국어 억양 생성

Stem-ML을 한국어에 적용하기 위해 크게 두 가지의 문제를 해결해야 한다. 첫째, 한국어의 경우 단어단위의 억양 패턴이 규칙화 되어 있지 않기 때문에 참조틀을 구성하기가 매우 어렵다. 영어의 경우 억양 사전(word tone dictionary)과 같은 참조틀 구성을 위한 규칙이 언어학적으로 존재하지만 국어는 이러한 규칙이 없다. 단지 말의 장단에 대한 규칙만이 존재하며, 이 역시 젊은 계층을 중심으로 점차 사라지는 추세에 있다[4]. 본 연구에서는 이러한 문제를 해결하기 위해 기존의 K-ToBI 기반의 운율생성 모델을 사용하였다.

둘째 Stem-ML은 중요도에 따라 각기 다른 강도값을 설정해야 한다. 실제 각 단어에 대한 강도값을 정확히 설정하기 위해서는 문장에 대한 의미적 분석과 함께 상황적 맥락 정보가 포함되어야 한다. 하지만 의미의 분석이나 상황적 맥락에

대한 분석은 본 연구의 범위를 벗어나는 것이다. 본 연구에서는 억양의 중요도를 예측하기 위해 이호영[4] 등이 사용하는 핵억양 개념을 사용하였다.

국어를 비롯한 대부분의 언어에서 화자가 전달하고자 하는 억양의미(화자의 감정과 태도, 화용론적 의미)의 대부분은 문장의 끝부분에 얹히는 억양패턴에 의해 전달된다. O'Conner와 Arnold[5]는 억양의미의 대부분을 전달하는 말마디 억양 끝부분의 억양패턴을 핵억양(nuclear tone)이라 불렀다. 여기서 말마디는 숨쉬기 단위를 의미한다[4]. 영어의 핵억양은 말마디(intonation phrase)의 마지막 액센트 음절과 뒤이어 나오는 음절들에 얹힌다. 영어의 핵억양은 대부분 둘 이상의 음절에 얹히지만 말마디의 마지막 음절이 액센트 음절일 경우에는 한 음절에만 얹힌다. 그러나 한국어에서는 말마디의 마지막 음절에 대부분의 억양의미를 전달하는 억양패턴이 얹힌다. 이호영[4][6]등에서 말마디의 마지막 음절에 얹히는 억양패턴을 핵억양이라 불렀다.

본 연구에서는 이러한 핵억양을 Stem-ML에 적용하여 보다 높은 강도값을 주어 억양을 생성하였다. 핵억양의 대표 참조틀은 벡터양자화(Vector Quantization(VQ))를 이용하여 제작하였으며, 예측 규칙 생성을 위해 CART tool을 이용하였다[7].

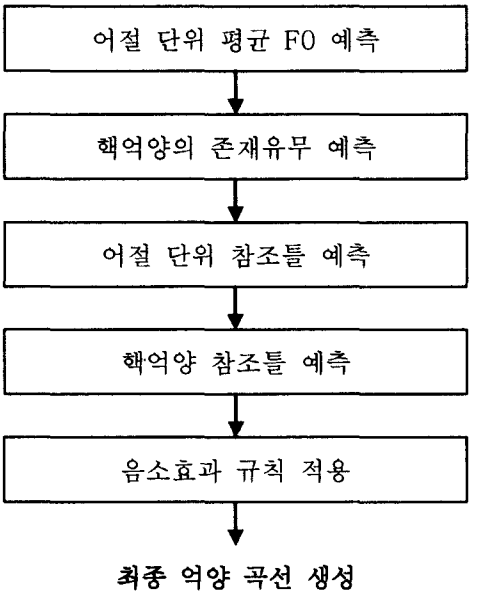
억양 생성의 단계는 <그림 2>와 같다. 이하에 핵억양 예측과 K-ToBI를 통한 어절 단위 예측을 중심으로 각각의 단계에 대해서 자세히 기술하도록 한다.

3.1. 전처리 단계

본 연구의 모델 생성을 위해 원광대학교 음성정보기술산업지원센터(SiTEC)에서 제작한 운율 합성용 음성 DB 중 음성문장 800개를 사용하였다. 이 DB의 음성파일은 16 kHz로 샘플링되어 16 bit로 양자화되었다. 녹음된 음성 코퍼스는 음소 단위와 어절단위로 레이블링되어 있다.

음성 데이터베이스로부터 Entropic 사의 ESPS/Xwaves+를 사용하여 10 ms 간격으로 기본주파수를 추출하였다. 원음의 기본 주파수 곡선은 성문(glottis)의 형태에 따라 발생하는 파형의 순간적 흐트러짐(jitter)이나 피치 흔들림(perturbation)의 영향으로 인하여 기본주파수 값이 갑자기 커지거나 작아지는 급격한 변화가 있을 수 있다. 이러한 급격한 변화는 억양의 모델을 생성할 때 악영향을 주게 됨으로 제거해야 한다. 또한 본 연구에서는 음소효과에 의한 기본 주파수 곡선의 변화는 독립적으로 가중합 할 것이기 때문에 원음성의 기본 주파수 곡선에서는 제거해야만 한다.

그러므로 전처리 단계는 1) 오류제거 단계(gross error correction), 2) 음소효과 제거단계 (microprosody removal), 3) 빈 구간에 대한 선형 보간법 (interpolation), 4) 부드러운 곡선 만드는 단계(smoothing) 등을 거치게 된다[8].



<그림 2> 억양 생성 순서도

3.2. 어절단위 F0 예측

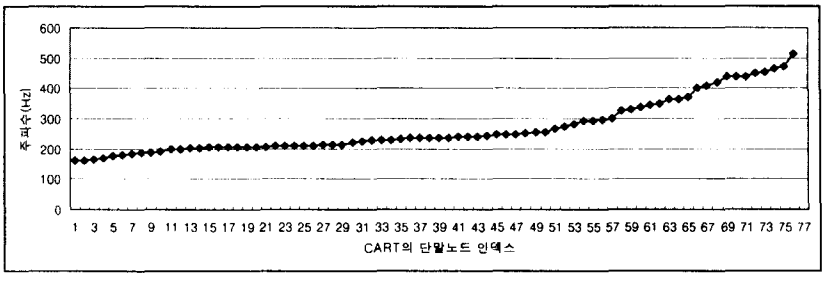
본 연구에서 앞서 설명했던 것처럼 억양구를 먼저 생성한 뒤, 이를 기반으로 하여 각 어절단위 및 핵억양의 참조틀을 이용하여 억양곡선을 생성하였다. 억양구 곡선을 생성하기 위해서 우선 각 어절단위의 평균 F0값이 얼마인지를 예측해야 한다. 이를 위해 CART를 이용하여 실변수 예측을 실시하였다. 예측에 사용된 변수는 <표 1>에 나타나있다. 훈련에 사용된 문장은 원광대 DB에 있는 문장 중 800문장을 사용하였으며, 테스트에 사용된 문장은 동일한 DB에서 추출한 200문장을 사용하였다. 예측 결과 <그림 3>과 같은 결과를 얻을 수 있었다.

<그림 3>은 어절단위 평균 F0의 예측결과를 보여준다. (a)는 각 단말노드의 주파수값을 나타내며 (b)는 실험 데이터들이 각 단말노드로 들어간 개수를 나타낸다. 대부분의 실험 데이터들이 기본주파수 150 Hz에서 240 Hz 사이에 분포하는 것을 알 수 있다. 실험에 사용된 화자의 평균 주파수가 183 Hz이었기 때문에 이러한 분포가 나타났다. (c)는 CART를 이용해 예측한 결과값과 실제값 사이의 표준 편차를 의미한다. 150 Hz에서 240 Hz 사이에 분포하는 단말노드의 값은 낮은 편차를 보인 반면, 그 이상의 값들에서는 높은 편차를 나타내었다.

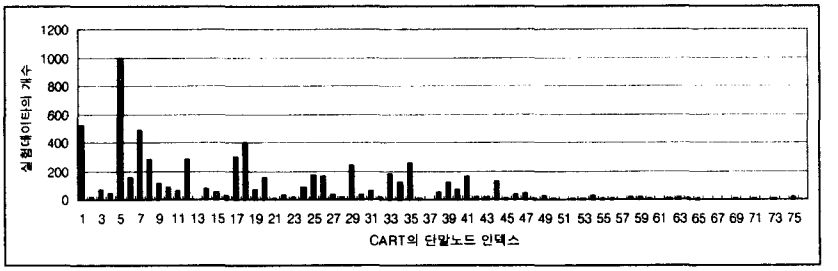
<표 1> CART 예측에 사용된 특징 변수

	변수 이름	설명
구분 변수	Bhpos	이전 어절의 첫 형태소 품사
	Btpos	이전 어절의 끝 형태소 품사
	Thpos	현재 어절의 첫 형태소 품사
	Ttpos	현재 어절의 끝 형태소 품사
	Nhpos	다음 어절의 첫 형태소 품사
	Ntpos	다음 어절의 끝 형태소 품사
	Tjosa	어절의 조사, 어미 유무
	OnsW	어절의 첫번째 음절의 초성
	OnsWM	어절의 첫번째 음절의 초성의 조음 방법
	Vow	어절의 첫번째 음절의 모음
	SylTW	어절의 첫번째 음절 유형
	DisGovernor	지배소까지의 어절 거리
	PhonW	어절의 음소 수
연속 변수	SylW	어절의 음절 수
	DisfirW	첫번째 어절까지의 어절 거리
	DisfirSy	첫번째 어절까지의 음절 거리
	DislastW	마지막 음절까지의 어절 거리
	DislastSy	마지막 음절까지의 음절 거리
	PreF0	앞 어절의 F0 값
	DisNucW	핵억양까지의 어절거리

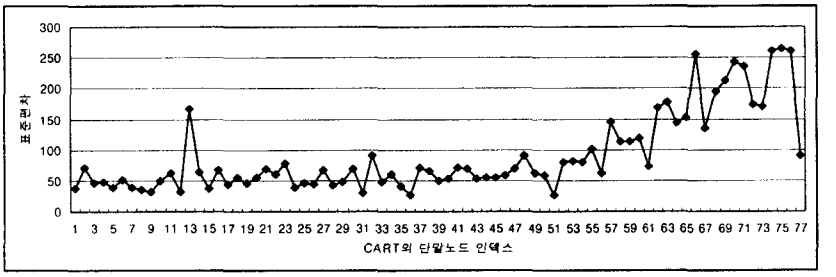
이는 훈련에 사용된 음성DB에 포함된 대부분의 어절단위 평균 F0가 150 Hz에서 240 Hz 사이에 분포하기 때문에 상대적으로 많은 데이터를 통해 정확한 규칙을 생성할 수 있었지만, 그 이상의 범위에서는 실제 훈련 DB에서 적게 분포하여 정확한 규칙을 생성할 수 없었기 때문이다. 150 Hz에서 240 Hz 사이의 표준편차는 58.1이고, 전체 데이터의 표준편차는 90.1로 얻어졌다.



(a)



(b)



(c)

<그림 3> 어절단위 평균 F0의 예측 결과
(a)각 단말노드의 주파수값, (b)각 단말노드로 입력된 데이터의 개수,
(c)각 단말노드의 예측값과 실제 테스트 데이터값과의 표준편차

3.3. 핵억양 존재 유무 예측

입력 문장을 어절별로 나누고, 어절의 끝에 핵억양이 존재하는지 예측한다. 예측은 CART를 통해 이루어졌으며 이 때 사용하는 특징변수도 <표 1>과 같다.

CART를 이용하여 원광대 음성 DB에서 추출한 800문장에서 핵억양의 존재유무를 위한 예측규칙을 생성하였다. 이를 훈련 문장 200문장을 이용하여 예측해 본 결과 <표 2>와 같은 결과를 얻었다. <표 2>에서 보는 바와 같이 핵억양의 예측 정확율은 84.5과 79.8로 상당히 높았다.

<표 2> 핵억양 존재 유무 예측 결과

실제값	경우의 수	무	유	예측 정확율(%)
무	4,983	4,211	772	84.5
유	1,865	374	1,482	79.8

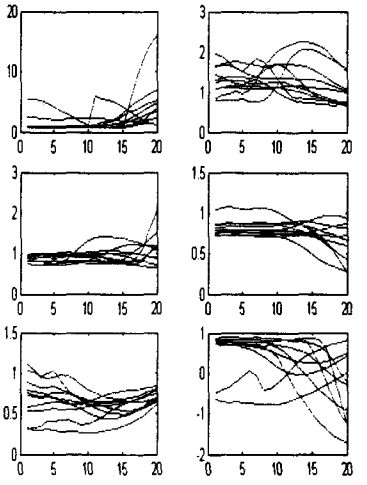
3.4. 핵억양 참조틀 예측

한국어에서는 핵억양이 어절의 가장 끝음절에 놓이는 것으로 알려져 있다. 본 연구에서 역시 한국어의 핵억양이 억양구의 가장 끝음절에 놓이는 것으로 가정하고 어절 단위 참조틀에서와 동일한 방법으로 VQ를 실시하였다. 이 때 VQ의 차수는 20차수로 정규화하였으며, codebook 크기는 64개로 하였다. VQ를 수행한 결과가 <그림 4>에 나타나 있다. 특정한 몇몇을 제외하고는 몇 개의 패턴으로 나누어질 수 있음으로 확인할 수 있다. 이중 가장 대표적인 패턴은 '수평조', '오름조', '내림조'이다.

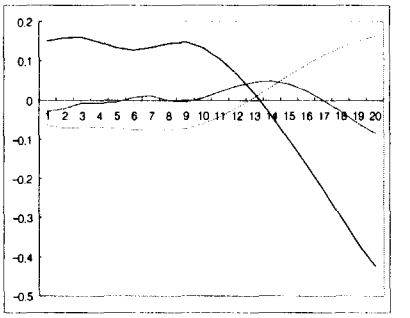
예측은 CART를 이용하였으며 이 때 사용한 특징변수는 <표 1>에서 PreF0를 제외하고 동일하게 사용하였다. 예측결과는 <표 3>에 나타나 있다.

<표 3> 핵억양의 참조틀 예측 결과

		경우의 수 (개수)	예측 결과(개수)			예측률(%)
			내림조	오름조	수평조	
실제	내림조	94	69	15	10	73.4
	오름조	107	33	62	12	57.9
	수평조	171	69	41	61	35.7



(a)



(b)

<그림 4> 핵억양의 억양 패턴
(a) VQ를 통한 전체 억양 패턴, (b) 핵억양 참조틀

3.5. K-ToBI 기반의 운율 모델을 이용한 어절단위 참조틀 예측

한국어는 영어와 달리 각 어절을 구성하는 단위가 단어가 아니라 어근과 어미, 혹은 단어와 조사의 결합형태가 대부분이다. 뿐만 아니라, 한국어는 영어에서와 달리 단어 자체가 가지는 강세(accent)가 존재하지 않으며, 단어의 억양 패턴은 문장에서 그 단어의 언어적 역할에 의해 결정된다. 한국어는 어절단위나 단어단위에 대한 참조틀을 구성하기 어렵다.

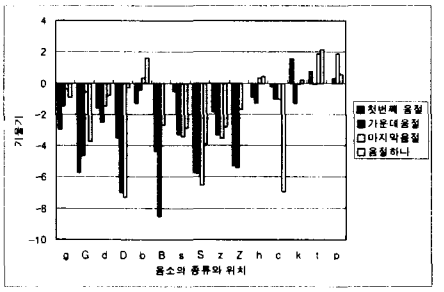
본 연구에서는 이러한 문제를 극복하기 위해 어절단위에 대한 참조틀을 기존에 본 연구실에서 개발된 K-ToBI 기반의 운율 생성 모델을 이용하였다. K-ToBI (Korean Tone and Break Indices)는 한국어에 대한 음조 표지와 끊어 읽기 정보를 분석하고 규칙화 하기 위한 전사 시스템이다[9]. 본 연구에 사용된 K-ToBI 기반의 운율생성 시스템은 장석복에 의해 제작된 시스템이다[10].

3.6. 음소효과의 적용

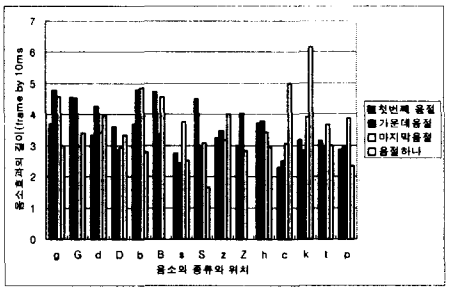
음소효과로 인해 선행하는 자음의 종류에 따라 후행하는 모음의 억양패턴은 다양하게 나타난다. /ㅍ, ㅅ, ㅋ, ㅊ/과 같은 무성자음 다음에 모음이 오면 억양곡선이 '하강음조'의 경향을 가지게 되고, /ㅂ, ㄷ, ㄱ, ㅈ/이 모음과 모음 사이에서 유성음화될 때 억양 곡선은 '상승음조'의 경향이 나타나게 된다. 본 연구에서는 음소효과에 따른 피치의 패턴변화를 억양 생성 모델에 효과적으로 적용하기 위해 원광대 DB에서 500문장을 추출하여 분석하였다.

우선 음소효과로 인해 후행하는 모음의 변화를 측정하기 위해 F0 변화값(delta F0)을 측정하였다. 구해진 F0 변화값이 특정 비교값(threshold) 보다 클 경우 음소효과에 의해 하강음조로 변경된 것으로 간주하였다. 또 구해진 F0 변화값이 특정 비교값보다 작을 경우 음소효과로 인해 상승음조로 변경된 것으로 간주하였다. 위 과정을 거쳐 음소효과가 있는 것으로 판단되는 프레임의 개수가 음소효과에 의해 영향을 받은 길이가 되며, 이 영역 안에 포함된 F0 변화값을 합하여 프레임 수로 나누면 음소효과를 표현할 기울기(slope) 정보가 된다.

<그림 5>와 <그림 6>은 음소효과로 인해 변경된 후행모음의 평균 F0 값과 효과를 받은 것으로 추정된 길이(frame 수)를 보여준다. 선행 자음이 경음일 경우 가장 높은 기울기를 가지는 것으로 나타났고, /ㅋ, ㅌ, ㅊ, ㅍ/ 등과 같은 무성 마찰음이나 파열음에서 오히려 낮은 기울기 값을 가지는 것으로 나타났다. 이는 /ㅋ, ㅌ, ㅍ, ㅊ/ 경우 단일음절의 일부에서만 급격한 변화를 주는 것이 아니라, 완만하지만 모음전체나 혹은 그 다음의 음절에 까지 많은 영향을 주기 때문인 것으로 판단된다.



<그림 5 > 음소효과로 인해 변경된 후행 모음의 F0 기울기



<그림 6 > 음소의 종류와 위치에 따른 음소효과의 길이

4. 억양 생성 실험 및 결과

4.1. 실험 방법

억양생성 성능평가는 크게 두 가지 방식을 통해 이루어졌다. 첫째, 객관적 비교를 위해 원음과의 RMSE(Root Mean Square Error)를 계산하였으며, 이를 본 연구실에서 이전에 구현했던 K-ToBI 기반의 운율생성 방식[10]에도 동일하게 적용하여, 두 시스템간의 성능을 비교하였다. 이 실험에는 원광대 DB에서 훈련에 사용하지 않은 80문장을 사용하였다.

둘째, 주관적 비교를 위해 기존 방식과 본 방식을 통해 생성된 두 가지 억양곡선을 TD-PSOLA(Time Domain Pitch Synchronous OverLap Add)를 이용해 원음에 동일하게 적용하여, 두 억양곡선 사이의 선호도를 평가하였다. 주관적 비교 실험에는 원광대 DB에서 훈련에 사용하지 않은 문장 10개를 추출하여 억양 곡선을 생성하였다. 실험참가자는 남자 14명, 여자 1명으로 총 15명의 대학생원생들이 참여하였으며, 이중 7명은 본 연구실의 연구원들이었으며, 나머지 8명은 비음성 전공자였다. 두 집단간 선호도에 대한 통계적 차이는 유의미하지 않았다.

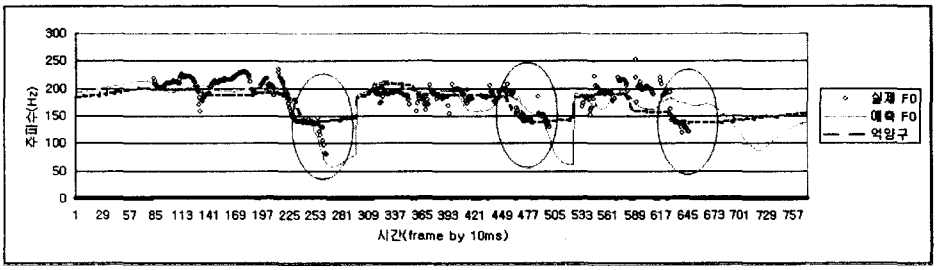
4.2 결과

원음과의 RMSE 비교에서 35.10(본 연구결과)과 36.85(기존 연구결과)로 본 연구의 결과가 약간 좋은 성능을 보였으나, 통계적으로 유의미한 차이는 없었다. 주관적 비교 실험에서는 54%(본 연구결과)와 46%(기존 연구결과)로 비교적 본 연구의 억양곡선 생성 방식이 우수한 것으로 나타났다.

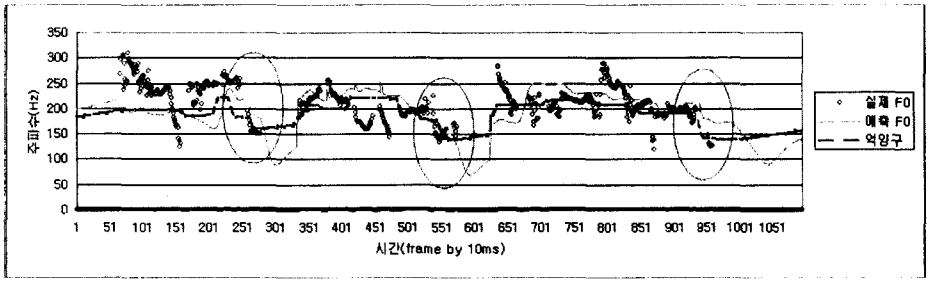
<그림 7>은 10개의 문장 중 가장 높은 선호도를 보인 억양 곡선과 가장 낮은 선호도를 보인 억양 곡선을 나타낸다. 각각 22.32와 43.88의 RMSE를 나타냈다. 두

그림 모두 둥근 원으로 표시된 영역이 핵억양 부분을 나타낸다. 핵억양 영역에서 원음의 패턴을 잘 모사하고 있음을 확인할 수 있다. 주관적 실험에 참여한 실험참가자들은 본 연구의 억양 생성기를 통해 생성된 합성음 중 말마디의 마지막 부분이 모두 안정적이라는 평가를 해주었으며, 전체적인 억양의 자연성에 큰 영향을 준다고 대답했다.

실험적 선호도에 대한 데이터와 실험참가자들의 응답을 바탕으로, 핵억양에 대한 독립적 모델링이 전체 억양의 자연성에 큰 도움을 준다고 판단할 수 있었다.



(a)



(b)

<그림 7> Stem-ML을 통한 억양 곡선
(a) 가장 높은 선호도를 나타낸 억양 곡선의 예
(b) 가장 낮은 선호도를 나타낸 억양 곡선의 예

5. Stem-ML을 이용한 정서음성 합성

TTS 시스템의 자연성과 명료도는 상업적으로 널리 이용될 만큼 큰 향상을 이루었다. 하지만 이는 인간 발화의 극히 일부분이라고 할 수 있는 낭독체에 국한된 것으로 인간이 가진 다양한 발화 방식을 모사하기에는 아직 많은 문제가 남아 있다. 그 중 대표적인 것이 인간의 정서적 표현을 어떻게 모사할 것인가 하는 문제이다. 많은 연구자들이 정서음성을 모사하기 위해 다양한 연구를 하였음에도 불구

하고 그 성과는 그리 만족할 만한 수준이 아니다. 인간의 정서는 실제 상황에서 아주 다양하게 나타나며, 동일한 정서상태일지라도 환경에 따른 차이와 개인에 따른 차이가 매우 크기 때문에 합성음 제작의 가장 기본이라고 할 수 있는 음성DB의 수집 자체가 매우 어려운 작업이다[11]. 더 나아가 수집된 정서음성을 이용해 특정 정서를 모델링하는 것은 다루어야 하는 음향적 파라미터의 수가 매우 많고 복잡할 뿐만 아니라, 사람에 따라 판단하는 정서의 유형에 크고 작은 차이가 있기 때문에 외적 타당도를 얻기가 쉽지 않다.

정서 자체에 대한 문제 역시 풀기 어려운 숙제로 남아 있다. 일단 정서가 무엇이고 이를 어떻게 구분해야 하는가 하는 문제와 특정 정서들 중 어느 것이 가장 전형적인 정서라고 할 수 있는가 하는 문제, 그리고 정서의 특징을 묘사하고 다른 정서와 구별되게 하며 모델링할 수 있게 하는 정서적 차원은 무엇인가 하는 문제 등이 남아있다. 하지만 이러한 많은 어려움에도 불구하고 다양한 형태의 합성방식을 이용한 정서음성 합성연구가 언어학과 심리학의 도움으로 여러 나라의 언어와 정서에 걸쳐 진행되어 왔으며 크고 작은 성과를 이루어왔다.

본 연구에서는 정서표현을 위한 여러 특징변수 중 가장 유력한 요소라고 할 수 있는 억양에 국한하여 정서음성을 모사하고자 하였으며 이는 앞서 설명한 Stem-ML 기반의 억양 모델을 통해 이루어졌다.

고정된 참조틀을 단순히 연결하거나 특정 목표점들을 구하여 이를 단순히 선형보간하는 것과 같은 기존의 억양 생성 모델에 비해 Stem-ML이 가지는 최대의 장점은 그 유연성에 있다. Stem-ML에서는 참조틀의 적용에서부터 억양과 관련된 대부분의 특징변수에 대해 각 태그값을 변경함으로써 언어학적 환경에 따라 각각 달리 적용할 수 있다. 각 태그값들은 억양 곡선을 위한 제약조건이 되고, 이들 제약조건들에 의한 최소자승해(least-square solution)를 이용하여 가장 부합하는 곡선을 그려낸다. 이러한 Stem-ML의 유연성이 정서음성과 같이 억양 패턴이 역동적으로 변화하는 발화에 대해 매우 적합할 것으로 판단하였다.

5.1. 억양생성 과정

정서음성 합성을 위해서는 무엇보다 그에 합당한 정서음성 데이터베이스가 마련되어야 한다. 국내에서는 극소수의 연구자들에 의해 소규모로 이러한 작업이 진행되고 있으며, 아직 코퍼스 기반의 정서억양 생성에 적합한 크기의 DB는 없는 실정이다.

충분한 규모의 정서음성 DB의 부재로 인해 Stem-ML에 필요한 참조틀 생성이 불가능하기 때문에, 본 연구에서는 기존의 낭독체 음성 DB에서 2000문장을 추출하여, VQ 과정을 거친 후 각 어절단위의 참조틀을 구성하였다. 정서 음성 합성에 사용된 어절 단위 참조틀은 <그림 8>과 같다.

<그림 8> 어절 단위의 참조틀

<그림 8>에 나타난 참조틀은 각 정서를 나타내기엔 매우 부족함에도 불구하고, Stem-ML이 가진 유연성을 통해 이를 극복할 수 있을 것인지 시도하고자 하였다. 또한 핵억양의 참조틀을 낭독체 억양 생성과 동일한 방법으로 적용하였다. 앞서 말한 바와 같이, 핵억양은 발화자의 정서, 태도 등을 표현하는 가장 중요한 억양 요소 중 하나이므로, 핵억양의 적절한 가공이 정서표현에 매우 중요할 것으로 판단하였다.

정서 음성의 억양 생성 역시 <그림 2>와 같은 낭독체 억양 생성의 순서를 그대로 따르도록 하였다. 다만, 낭독체 억양 생성의 대부분의 모듈에서 충분한 DB를 통해 예측 룰을 생성할 수 있었지만, 정서음성의 경우 이를 위한 충분한 DB가 부재하기 때문에, 모든 과정을 수동으로 적용하였다. 또한 "정말 그렇단 말이야?"라는 비교적 짧은 한 문장에 대해서만 이를 적용하였다.

적용한 정서는 비교적 가장 전형적인 정서로 알려져 있는 기쁨, 슬픔, 화냄, 지루함, 그리고 공포 등 5가지 정서[12]에 중립적 정서를 포함 총 6가지 정서이다.

실제로 정서음성 합성의 자연성은 억양이라는 단일 요소에 의해 결정되지는 않는다. 억양 이외에도 에너지, 지속시간, 끊어읽기의 빈도, 포먼트의 모양, 발음의 정확성 등과 같은 요소에 의해 많이 영향을 받는다. 하지만 본 연구는 억양을 중심으로 한 운율요소만을 통해 정서음성을 생성하였다. 이는 본 연구가 Stem-ML이 정서음성합성을 위한 적합한 억양 모델인지를 검증하는데 목적이 있기 때문에, 운율 이외의 다른 요소가 포함될 경우 억양에 의한 정서모델링의 적합성을 검증하는데 어려움이 있다고 판단했기 때문이다.

본 연구에 적용한 억양 이외의 운율요소로 에너지와 지속시간이며, 이는 모두 상수로 적용함으로써 적용효과를 최소화하였다. 또한 각 정서음성의 평균 F0를 적용하였다. 적용을 위한 변수값은 [12]에 포함된 정서음성들의 평균값을 취하였다.

Stem-ML에서 가용한 변수는 일반적으로 40개를 넘는다. 하지만 이렇게 많은 변수를 각 정서음성에 맞게 수동으로 조율한다는 것은 매우 어렵다. 본 연구에서는 실험자가 가장 중요한 요소로 판단한 6개의 태그에 대해서만 이를 조율하였다.

태그에 대한 자세한 사항은 [1]에서 찾을 수 있다. 태그의 수정값은 <표 4>에 나타나 있다.

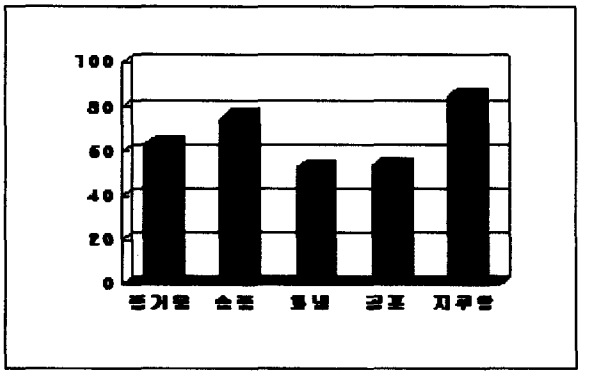
5.2. 실험 및 결과

정서음성의 자연성을 평가하기 위해 10명의 실험참가자가 참여하였다. 모든 실험참가자는 부산대학교 음성통신연구실의 연구원들로 구성되었다.

실험은 특정 정서로 가공된 정서음성을 들려주고 그 정서가 어떤 정서인지를 판단하게 하는 정서판별실험으로 이루어졌다. 실험참가자는 한 문장을 청취한 6개의 정서 중 하나를 선택하게 된다. 이때, 한 정서가 여러 개로 표현될 수 있음을 알려줌으로써 선택지에 의해 발생할 수 있는 편향(bias)을 줄이도록 하였다.

<표 4 > 정서별 태그 수정값

	Set smooth	Set pdroop	Set base	Set adroop	Step strength	Stress type
즐거움	0.06	0.001	244	8.3	1.5	1.0
슬픔	0.06	0.002	150	2.5	0.5	0.2
화냄	0.06	0.001	239	2.5	1.8	1.0
지루함	0.06	0.002	202	10.0	0.2	0.1
공포	0.00	0.0001	248	1.0	2.0	1.0



<그림 9> 정서판별 결과

이를 통해 생성된 정서음성의 판별실험 결과가 <그림 9>에 나타나 있다. 그림에서 보는 바와 같이 기쁨과 슬픔, 지루함 등과 단조로운 정서음성의 경우 각각 65%, 75%, 85%를 비교적 높은 판별율을 보였다. 하지만, 화냄과 공포와 같이 억양의 변화가 역동적으로 변하는 정서의 경우, 54%와 53%로 낮은 정서 판별율을 나타내었다.

정서음성 데이터베이스의 부족으로 각 정서음서에 대한 억양 패턴의 참조틀을 제작하지 않은 상태에서 실시한 본 예비실험의 결과를 바탕으로, Stem-ML의 유연성이 정서음성의 미묘한 변화를 모사할 수 있다는 가능성은 충분히 확인된 것으로 판단된다.

6. 결론

Stem-ML은 인간의 발화적 특성에 기반한 억양 모델로서 각 태그값들을 변경 적용함으로써 다양한 발화적 특성을 표현할 수 있는 특징을 가지고 있다. 본 논문에서는 이러한 Stem-ML을 기반으로 한국어 억양 생성 방식을 구현하고, 이를 바탕으로 정서음성 합성을 위한 억양 곡선 생성 가능성을 검토하였다.

한국어는 단어 단위의 억양패턴이 규칙화되어 있지 않기 때문에 어절 단위의 억양패턴은 기존의 K-ToBI 기반의 억양 생성 모델을 이용하여 찾아내었다. 핵억양의 존재유무, 어절의 평균 F0, 어절단위 및 핵억양의 참조틀 예측은 모두 CART를 이용해 규칙을 생성하여 실시했다. 음소효과의 영향을 반영하기 위해 원광대 DB에서 500문장을 분석해 규칙을 생성하여 적용하였다. 실험결과 기존의 억양생성 방식과의 비교에서 우수한 성능을 냈으며, 특히 핵억양의 모사에서 매우 높은 성능을 보였다. 이를 통해 Stem-ML이 한국어 억양 생성에도 효과적인 모델임을 확인할 수 있었다.

Stem-ML의 한국어 적용과 함께, 본 연구에서는 즐거움, 슬픔, 화냄, 지루함, 공포 등의 다섯 가지 정서음성을 합성하였다. Stem-ML에서 사용하는 태그값들 중 일부를 선별하여 적용함과 동시에 핵억양의 평균 피치와 핵억양의 참조틀을 일부 수정함으로써 즐거움, 슬픔, 지루함과 같이 억양변화가 비교적 작고 단조로운 정서에서 자연스런 억양곡선을 생성할 수 있었다. 하지만 화냄, 공포와 같이 억양변화가 크고 끊김이 자주 들어가는 정서는 단순히 태그값을 변경하는 것만으로는 적합한 억양을 생성할 수 없었다. 여기에 억양 생성시 각 정서에 맞는 참조틀을 생성하지 못하였기 때문에 부자연성은 더욱 컸다. 적절한 정서음성의 합성을 위해서는 무엇보다 충분한 정서음성 데이터베이스가 마련되어야 할 것으로 판단된다.

참 고 문 헌

[1] G. Kochanski and C. Shilh, "Prosody modeling with soft templates", *Speech Communication*, Vol. 39, pp. 311-352, 2003.

[2] G. Kochanski and C. Shilh, "Automatic modeling of Chinese intonation in continuous speech", *in Proc. of EUROSPEECH*, Aalborg, Denmark, pp. 911-914, 2001.

[3] G. Kochanski and C. Shilh, "Hierarchical structure and word strength prediction in Mandarin prosody", *in Proc. of 4th ISCA Workshop on Speech Synthesis*, pp. 217-222, Sep. 2001.

[4] 이호영, 국어 음성학, 태학사, 1996년.

[5] J. D. O'Connor and G. F. Arnold, *Intonation of Colloquial English*, London: Longman, 1973.

[6] 이호영, 국어 운율론, 한국연구원, 1997년.

[7] *Classification and Regression Tree(CART)*, Salford Systems, 2002.

[8] H. Fujisaki, S. Narusawa and M. Maruno, "Pre-processing of fundamental frequency contours of speech for automatic parameter extraction", *in Proc. of ICSP*, pp. 722-725, 2000.

[9] M. E. Beckman and S. A. Jun, "K-ToBI(Korean ToBI) labeling convention version 2.1", Revised, Nov. 1996.

[10] 장석복, TTS를 위한 자동 억양곡선 생성방식에 관한 연구, 부산대학교 인지과학협동과정 석사학위논문, 1999년 2월.

[11] 조철우, 김대현, "멀티미디어 환경을 위한 정서음성의 모델링 및 합성에 관한 연구", 한국음성과학회 Vol. 5, No.1. pp. 35-47, 1999.

[12] 음성정보기술산업지원센터, 정서음성 데이터베이스 Emotion 01. 2004년.

접수일자: 2005년 5월 15일
게재결정: 2005년 6월 21일

▶ 한영호(Young Ho Han)
주소: 137-792 서울시 서초구 우면동 17번지
소속: KT 마케팅 연구소
전화: 02) 526-6786
E-mail: yhhan@kt.co.kr

▶ 김형순(Hyung Soon Kim)
주소: 609-735 부산광역시 금정구 장전동 산 30번지
소속: 부산대학교 전자공학과
전화: 051) 510-2452
E-mail: kimhs@pusan.ac.kr