한국어 음성합성기의 성능 향상을 위한 합성 단위의 유무성음 분리*

Separation of Voiced Sounds and Unvoiced Sounds for Corpus-based Korean Text-To-Speech

홍 문 기** · 신 지 영*** · 강 선 미**
Munki Hong · Jiyoung Shin · SunMee Kang

ABSTRACT

Predicting the right prosodic elements is a key factor in improving the quality of synthesized speech. Prosodic elements include break, pitch, duration and loudness. Pitch, which is realized by Fundamental Frequency (F0), is the most important element relating to the quality of the synthesized speech. However, the previous method for predicting the F0 appears to reveal some problems. If voiced and unvoiced sounds are not correctly classified, it results in wrong prediction of pitch, wrong unit of triphone in synthesizing the voiced and unvoiced sounds, and the sound of click or vibration. This kind of feature is usual in the case of the transformation from the voiced sound to the unvoiced sound or from the unvoiced sound to the voiced sound. Such problem is not resolved by the method of grammar, and it much influences the synthesized sound. Therefore, to steadily acquire the correct value of pitch, in this paper we propose a new model for predicting and classifying the voiced and unvoiced sounds using the CART tool.

Keywords: Speech Synthesize, Pitch, Fundamental Frequency, Unit of Triphone, CART

1. 서 론

최근 음성 인식 및 음성 합성 기술을 이용한 제품에 대한 수요가 늘어나면서, 한국어 음성 처리 성능 향상을 목적으로 한 연구가 다각도로 이루어지고 있다. 특히, 음성 합성 관련 분야는 웹사이트는 물론이고 ARS(Automatic Response Service), 음성 메일, 기타 음성 관련 제품들을 통하여 널리 서비스되고 있으나, 청취자들은 좀더 자연스러운 음을 요구하고 있다. 그러므로 좀더 자연스러운 음을 생성하는 합성기의 개발은 주요한 연구 주제이다.

일반적으로 코퍼스 기반 음성 합성에서는 합성 단위별로 예측값을 후보 단위 중에서 가져오는 방법을 이용하여 음성을 합성하는데, 이 때 유무성으로 분류된 음성의 실제 실현이 규정된 것과 다르면, 합성음이 울리는 소리가 나고 부자연스럽게 된다[1]. 본 논문에서는 피치 값

* 본 논문은 한국과학재단 목적기초연구(R01-1999-000-00229-0) 지원으로 수행되었음.
** 서경대학교 컴퓨터과학과
*** 고려대학교 국어국문학과

을 정확히 구하기 위한 목적으로 유무성음을 정밀하게 분리하기 위해 CART (Classification and Regression Tree)[2]를 이용한 새로운 모델을 제안하였다.

본 논문에서는 다음과 같이 연구를 진행하였다. 첫째, 기존의 CART 툴을 사용한 연구들은 음소 피치나 어절 피치를 예측하는 데에 머물렀다. 유무성음 여부에 대한 부정확한 예측(경음이나 격음에 대한 문법적인 분리)은 음소 피치 평균은 물론, 어절 피치 평균값에 영향을 미치게 된다. 따라서 유무성음을 정확하게 구분하기 위해 음성 DB(수동 레이블링을 거친 DB)와 문법을 사용하여 만든 파라미터(유무성음 구분)를 CART DB로 사용하여 정확한 유무성음 분리 실험을 하고, 그 결과를 문법적 규칙을 이용한 분리 확률과 비교하여 평가하였다. 둘째, 한국어의 경우 대개 유성음은 중성에 무성음은 종성에서 위치하며, 유무성음 구분이 가장 어려운 곳은 초성이다. 하지만 실제 음성 데이터에는 종종 중성에도 무성모음화로 인하여 무성음이 존재하기도 하고 유성음화로 인하여 종성에도 유성음이 나타나곤 한다. 초성을 실험하면 대부분의 유무성음을 분리해 낼 수 있지만, 종성에도 유무성음을 가지고 있으므로 각 경우에 대하여 실험을 실시하여 규칙을 얻은 뒤 예측 결과치와 비교한다.

2. 본 론

음성 합성기는 기능별로 언어 처리부, 운율 생성부, 음성 합성부로 나뉜다. 본 논문은 운율 생성부에 해당하는 피치 예측을 위해 유무성음을 분리하기 위한 방법에 관한 것이며, 초성, 중성, 종성 각각의 경우에 대해서 실험하였다. 실험에 사용된 자료는 본 연구를 위해서 언어학 전문가가 선정한 800 개의 문장에서 추출한 약 65,000 개의 음소였다. 음성 합성기에 사용되는 운율 요소들을 예측하는 데 일반적으로 사용하는 방법으로는 HMM이나 CART를 사용한 모델링 방법[3]이 있다. 본 논문에서는 확률적으로 더 나은 결과[1]를 기대할 수 있는 CART를 사용하여 모델링을 하였다. 표 1은 음성학 전문가로부터 얻은 수동 레이블링 데이터의 파라미터를 나타낸다.

표 1에 보인 파라미터들을 사용하여 CART의 DB를 구축하였으며, 사용된 파라미터 중 IPHONE 파라미터의 정보는 표 2와 같다[4]. 유무성음 구분은 문법적인 절차를 거쳐 분리된 정보를 사용하였으며, 이는 바로 DVOICE 파라미터의 정보이기도 하다. 모두 45 개의 음소가 있으나 DB 구축에 편의상 허음소(dummy phoneme)를 7 번과 40 번에 넣어 사용하였다. 7 번은 'o'의 초성으로서 실제 존재하지 않으며, 40 번은 중성과 종성 간에 구분 인덱스로 사용하였다.

표 1. 음소 피치 예측을 위해 사용된 파라미터(labeling data)

파라미터 종류	내 용
DVOICE	관측 음소의 문법적 유무성음 구분
DDUR	관측 음소의 지속시간
IPHONE	관측 음소
APPOS	AP 내에서의 음소 위치
IAP1	관측 음소 앞의 IP(Intonation Phrase), AP(Accentual Phrase) 유무
IPPIT	관측 음소의 앞 음소 피치
IRPHONE	관측 음소의 뒤 음소
DLOCPCT	문장 내에서의 관측 음소 위치(백분율)
ILPHONE	관측 음소의 앞 음소
IPPOS	IP 내에서의 관측 음소 위치
ILOC	문장 내에서의 관측 음소 위치

표 2. IPHONE 파라미터 정보

음소번호	초성	유무성음구분	음소번호	중성	유무성음구분	음소번호	종성	유무성음구분
1	g	유	20	a	유	40		
2	d	유	21	e	유	41	gf	무
3	b	유	22	E	유	42	df	무
4	n	유	23	i	유	43	bf	무
5	m	유	24	o	유	44	nf	유
6	l	유	25	u	유	45	mf	유
7			26	U	유	46	lf	유
8	s	무	27	v	유	47	Nf	유
9	z	유	28	wa	유			
10	h	유	29	we	유			
11	G	무	30	wE	유			
12	D	무	31	wi	유			
13	B	무	32	wv	유			
14	S	무	33	ja	유			
15	Z	무	34	je	유			
16	k	무	35	jE	유			
17	t	무	36	jo	유			
18	p	무	37	ju	유			
19	c	무	38	jv	유			
			39	xi	유			

표 2에 보인 파라미터를 사용하여 결정 트리를 구성하였다. 이 실험은 초성, 중성, 종성을 나누어 각각 모델을 만들고 그 결과를 바탕으로 비교 분석하는 방법으로 수행되었다.

2.1 CART를 이용한 피치 예측 실험

피치는 성대의 진동수와 관련된 것으로 성별, 연령 등에 따라 달라지며, 또한 화자의 개인적인 특징을 나타내는 요소이기도 하다. 음성 합성에서 피치는 다른 운율적 요소들, 즉 길이, 강세보다 큰 가중치가 부여되는 것이 일반적이다. 음성합성 시에는 문장, 어절, 음소에 대한 피치값을 구하는 것이 보통이지만, 본 실험에서는 음소에 대한 피치값을 구해보고, 가중치의 변화에 따른 합성음의 음질의 변화를 알아보도록 하겠다.

주어진 DB를 유성음과 무성음으로 분리하고 유성음은 초성, 중성, 종성으로 나누어 실험하였고, 문법적으로 무성음으로 나타나는 음소는 실험에서 제외하였다.

실험 1은 무성음 음소를 제외한 음소 피치 예측과 관련된 것이다. 음소 피치 예측을 위해 사용한 회귀 트리 모델링[3]에서 Splitting Method는 Least Squares[1]을 선택하였고, SE (Standard Error) Rule은 최소 비용 트리를 선택하였다. 10-fold cross-validation 방법(학습 데이터의 10분의 1을 테스트 데이터로 사용하는 방법)에 의해 최적의 트리를 결정하였다.

표 3은 모델을 구성하는 데 기여한 파라미터의 중요도를 나타낸 것이다. 표에 보인 바와 같이 중성이나 종성은 좌, 우 경계 피치에 의해, 초성은 문법적인 유무성음 정보에 의해 가장 많은 영향을 받는다. 그 이유는 중성은 거의 모든 음소가 유성음, 종성은 무성음이 대부분으로 유무성음이 이렇게 확고한데 비해 초성은 유성음화를 겪는 음소가 많기 때문이다. 유무성음의 결정에 따라 피치예측이 결정되는 것은 당연한 결과라고 볼 수 있다.

예측 결과 표 4에서 보인바와 같이 초성, 중성, 종성 별로 평균 피치를 얻을 수 있다. 표 4에서 나타난 무성음은 문법적으로 걸러지지 않는 음소로서 실제 DB를 체크하여 피치가 0인 것을 얻은 것이다. 중성은 모두 모음이기에 피치값이 217.39로 가장 높은 것을 알 수가 있었다. 반면, 초성(문법적으로 분리되어 격음과 경음이 빠진 유성음 음소: IPHONE 1~6, 9~10)은 평균이 121.92로 나왔다. 무성음을 제외한 초성 음소를 평균 낸 것인데 기대했던 것 보다 낮게 나온 것을 알 수 있다. 이는 유성음인 음소들 가운데 앞뒤의 음소에 영향을 받거나 해당 음소의 지속시간이 짧아 앞뒤 무성음의 영향을 받아 피치가 0이거나 0에 가까운 값이 되어 무성음화되는 현상으로 이런 일부 음소들로 인해 평균 피치에 지대한 영향을 미친다. 그뿐만 아니라 합성 시에 트라이폰 후보 중 가장 적당한 후보를 선택하여 합성하게 되므로 무성음화된 음소를 유성음으로 오인하여 합성하게 되면 합성 결과가 틀려지게 된다.

표 3. CART에 의한 음소 피치 예측 변수 중요도(초성, 중성, 종성)

초 성		중 성		종 성	
Variable	기여도(백분율)	Variable	기여도(백분율)	Variable	기여도(백분율)
DVOICE	100.00	IRPIT	100.00	ILPIT	100.00
DDUR	65.68	DDUR	31.96	IRPIT	61.32
APPOS	63.58	ILPIT	26.90	IPPIT	41.89
IAP1	62.46	IRPHONE	20.56	IRPHONE	20.28
IRPIT	58.08	IP	20.13	IAP2	19.93
ILPIT	54.76	IAP2	18.36	DLOCPCT	11.95
DEOJPOS	41.45	IPPOS	15.98	DLAST	8.39
DEOJPCT	26.05	DLOCPCT	15.40	IPPOS	7.13

표 4. 음소당 평균 피치와 유무성음 비율

음 소	평균 피치	실제 DB 평균 피치	실제 DB 무성음 비율 (음소개수 : 무성음)
초 성	121.92	204.23	17,573 : 7,082
중 성	217.39	220.19	30,168 : 383
종 성	214.00	216.00	8,838 : 48

표 4에 보인 바와 같이 실험결과에 따르면 유성음과 무성음을 구분하는 것이 가장 필요한 부분은 초성임을 알 수 있다. 초성의 경우 무성음이 자주 나타날 뿐 아니라, 평균 피치의 차이도 약 1.6 배의 차이가 발생한다.(121.92 Hz 대 204.23 Hz) 이런 무성음을 찾아내어 유성음에서 분리하기 위해 실험 2를 통해 추가적인 모델을 만들어 유무성음을 분리해 내고자 한다.

2.2 CART를 이용한 유무성음 분리 실험: 초성

실험 1을 통해 유무성음의 분리가 정확한 피치 예측에 매우 중요하다는 것을 알게 되었으며, 본 실험에서는 문법적인 정보만으로는 분리할 수 없고, 환경에 가장 많은 영향을 받는 초성을 가지고 실험하기로 한다. 분리 규칙을 생성하기 위해 분류트리(Classification Tree)를 선택하였다.

실험 2는 초성의 유성음 중 무성음을 예측하는 것이다. 음소 피치 예측을 위해 사용한 분류트리 모델링[3]에서 Splitting Method는 Gini Index[1]을 선택하였고, SE(Standard Error) Rule은 최소 비용 트리를 선택하였다. 10-fold cross-validation 방법에 의해 최적의 트리를 결정하였다.

실험 2에 사용된 파라미터의 기여도는 표 5와 같다.

표 5. CART에 의한 무성음화 음소 분리 예측 변수 중요도(초성)

Variable	기여도 (백분율)
DVOICE	100.00
DDUR	73.82
IPHONE	71.02
APPOS	20.87
IAP1	19.63
IPPIT	9.67
IRPHONE	2.62
DLOCPCT	2.15
ILPHONE	1.03
IPPOS	0.90
ILOC	0.62

실험 결과 DVOICE, DDUR, IPHONE 파라미터 정보가 차례대로 기여도를 나타내고 있었

다. 하지만 DVOICE나 IPHONE은 문법적으로 얻을 수 있는 파라미터이므로 이번 모델링에 효과적인 기여도를 나타내는 것은 DDUR라고 할 수 있겠다. 이는 음소의 길이에 따라 해당 음소와 유성음과 무성음을 구분할 수 있는 척도가 된다고 할 수 있다. 또한 다음 표 6을 보면 본 논문에서 제안한 실험 2에 의해 제거된 무성음화 음소들의 비율을 알 수 있다.

표 6. 실험 2, 유무성음 분리의 예측 확률(학습 데이타)

유무성음 구분	전체 개수	유무성음 분류율	무성음	유성음
무성음	7,082	89.090	6,309	773
유성음	10,491	87.504	1,311	9,180

표 7. 실험 2, 유무성음 분리의 예측 확률(테스트 데이타)

유무성음 구분	전체 개수	유무성음 분류율	무성음	유성음
무성음	7,082	86.782	6,145	936
유성음	10,491	87.017	1,362	9,129

실험에 사용된 데이터 초성 17,573 개의 음소는 7,082 개의 무성음화된 음소와 10,491 개의 유성음 데이터로 구성되어 있다. 실험 2를 통해서 테스트 데이터의 경우 무성음화된 음소를 예측한 결과 7,507 개를 찾을 수 있었다. 그 중에는 유성음을 무성음으로 오분류한 1,362 개의 음소는 피치 예측에 활용되지 못했다. 그러나 전체 음소 17,573 개 중에 6,145 개의 무성음을 찾아서 제거함으로 인해 정확도가 더 높은 피치예측을 할 수 있게 되었다.

3. 결 론

무성음 음소를 제외한 음소 피치 예측 실험(실험 1) 결과를 살펴보면, 초성의 경우 문법적인 유무성음 분리를 거쳤음에도 불구하고 무성음의 분포가 크게 나타나기 때문에 초성에 대한 유무성음 분리가 불가피하다는 결론을 내릴 수 있었다. 실제 DB상에서 무성음화된 7,082 개를 찾기 위해 실험 2를 실시한 결과 해당음소나 주변 음소 환경에 따라 유성음을 구분하는 모델을 만들어 낼 수 있었다.

다음은 실험 2를 통해 분리된 결과를 실험 1에 재적용하여 평균 피치를 다시 예측한 결과이다.

전체 초성음의 유성음 개수	= 17,573
(문법적으로 분리된 유성음 개수)	
무성음화된 음소 개수	= 7,507
유성음 개수	= 10,066

무성음 분리 전 음소 평균 피치[실험 1] = 121.92
유무성음 실험 후 음소 평균 피치[실험 2] = 192.43
실제 DB상의 평균 피치 = 204.23

상기 실험 결과에서 알 수 있듯이 실험 1에서는 평균 피치값을 121.92로 얻었지만, 실험 2를 통해 192.43으로 목표 값 204.23에 더 가까워졌다. 그러나 평균 피치만 보고 예측 값이 정확히 볼 수 없으므로 RMSE[1] 측정값을 사용해 표 8과 같이 비교하였다. 표 8의 모델 적용 전 값은 실제 DB값과 실험 1을 통해 얻은 값의 RMSE를 구한 것이며, 모델 적용후의 값은 실제 DB값과 실험 2를 통해 얻은 값의 RMSE를 구한 것이다.

표 9. RMSE 측정값으로 비교

	모델 적용 전	모델 적용 후
초성	46.917777	17.743544
중성	6.957011	68.387864
종성	0.534522	6.513722

표 8을 보면 알 수 있듯이 유성음의 무성음화가 많은 초성의 경우 본 실험을 통해 얻은 모델이 효과가 있음을 알 수 있다. 한국어 합성기에서 합성 단위를 찾아 낼 때 초성의 유무성음을 본 논문에서 제안한 방법을 통하여 분리한 후 평균 피치를 예측하면 좀 더 정확한 피치 값을 예측할 수 있으며, 따라서 이 값을 이용하여 합성을 할 경우 합성음의 음질 향상을 기대할 수 있을 것이다[5]. 단, 본 논문에서 제안한 방법에 의해 합성된 음의 자연성에 대한 청취 실험은 아직 실시하지 못하였으므로 이는 향후의 연구과제가 될 것이다. 또한, CART툴을 사용하여 정확률을 높일 수 있는 방법들이 제안되어 있어[6] 좀더 향상된 정확성을 기대할 수 있다.

참 고 문 헌

[1] 박상언. 2001. *코퍼스 기반 한국어 음성합성 시스템의 합성음 자연성 향상*. 전남대학교 대학원 석사논문.
[2] *An Overview of the CART Methodology*. http://www.salford-systems.com
[3] 이상호. 1999. *한국어 TTS시스템을 위한 운율의 트리기반 모델링*. KAIST 석사논문.
[4] 문화관광부 21세기 세종 계획. 2000. "국어 기초자료 구축." p. 188
[5] 김병창, 이진석, 이근배, 권오일. 2000. "무제한 단어 한국어 TTS 시스템 구현." *제17회 한국음향학회 학술대회 논문집*, p. 298.
[6] Breiman, L. 1996. "Bagging predictors." *Machine Learning*, 24, 123-140.

접수일자: 2003. 4. 24.
게재결정: 2003. 5. 28.

▲ 홍문기
서울특별시 성북구 정릉 4동 (우: 136-704)
서경대학교 컴퓨터과학과
Tel: +82-2-940-7291 Fax: +82-2-919-5075
E-mail: chatmunk@daum.net

▲ 신지영
서울특별시 성북구 안암동 5가1번지 (우: 136-701)
고려대학교 국어국문학과
Tel: +82-2-3290-1973 Fax: +82-2-925-2506
E-mail: shinjy@korea.ac.kr

▲ 강선미
서울특별시 성북구 정릉 4동 (우: 136-704)
서경대학교 컴퓨터과학과
Tel: +82-2-940-7291 Fax: +82-2-919-5075
E-mail: smkang@skuniv.ac.kr