Text-to-Speech 변환 시스템을 위한 회귀 트리 기반의 음소 지속 시간 모델링

표경란, 김형순
{knpyo, kimhs}@hyowon.pusan.ac.kr
부산대학교 인지과학협동과정, 부산대학교 전자공학과

Regression Tree based Modeling of Segmental Durations For Text-to-Speech Conversion System

Kyung Ran Pyo, Hyung Soon Kim
Dept. of Interdisciplinary Research Program of Cognitive Science,
Graduate School, Pusan National University,
Dept. of Electronics Engineering, Pusan National University

요 약

자연스럽고 명료한 한국어 Text-to-Speech 변환 시스템을 위해서 음소의 지속 시간을 제어하는 일은 매우 중요하다. 음소의 지속 시간은 여러 가지 문맥 정보에 의해서 변화하므로 제어 규칙에 의존하기 보다 방대한 데이터베이스를 이용하여 통계적인 기법으로 음소의 지속 시간에 변화를 주는 요인을 찾아내려고 하는 것이 지금의 추세이다. 본 연구에서도 트리 기반 모델링 방법중의 하나인 CART(classification and regression tree) 방법을 사용하여 회귀 트리를 생성하고, 생성된 트리에 기반하여 음소의 지속 시간 예측 모델과, 자연스러운 끊어 읽기를 위한 휴지 기간 예측 모델을 제안하고 있다. 실험에 사용한 음성코퍼스는 550개의 문장으로 구성되어 있으며, 이 중 428개 문장으로 회귀 트리를 학습시켰고, 나머지 122개의 문장으로 실험하였다. 모델의 평가를 위해서 실제값과 예측값과의 상관관계를 구하였더니 음소의 지속 시간을 예측하는 회귀 트리에서는 상관계수가 0.84로 계산되었고, 끊어 읽는 경계에서의 휴지 기간을 예측하는 회귀 트리에서는 상관계수가 0.63으로 나타났다.

1. 서론

컴퓨터와 신호처리 기술의 급속한 발전에 따라, 음성은 사람과 사람 사이의 의사전달 수단으로서 뿐만 아니라 인간과 컴퓨터 사이의 의사 소통을 위한 매개체로서의 역할이 요구되기에 이르렀고 실제로 사용되기 시작하고 있다. 문장-음성 변환(Text-to-Speech(TTS) Conversion) 기술은 이러한 요구에 부합되는 것으로 사용자가 음성으로 바꾸고자 하는 문장을 기계에 입력하면, 기계가 그 문장을 분석한 후 미리 저장된 말소리의 기본 단위들로부터 사용자가 원하는 음성을 신호로 합성하는 것을 말한다[1]. 이때 어떤 TTS 합성 시스템이라도 그 하위 모듈로서 음소의 지속 시간을 예측하는 모델을 가지게 되는데, 합성음을 자연스럽고 명료하게

생성하기 위해서는 최대한 자연음에 가깝게 생성해야 한다[2]. 그러나 음소의 지속 시간은 여러 가지 문맥 정보에 의해서 변화하게 되므로 모든 상황을 고려하여 규칙을 세우는 데는 무리가 있다. 따라서 방대한 음성 데이터베이스에서 통계적인 방법으로 음소의 지속 시간에 변화를 미치는 요인을 찾아내고 이를 음소 지속 시간 모델의 제어 요인으로 사용하려는 시도가 많이 이루어 지고 있다. 트리 기반 모델링 방법중의 하나인 CART 방법은 특징값으로 카테고리변수와 실변수를 동일한 시각에서 모두 처리할 수 있다는 것과, 그 결과를 결정 트리나 회귀 트리로 만들 수 있어 쉽게 해석이 가능한 장점이 있어 음소 지속 시간에 대한 연구에서 많이 사용되고 있다[3].

본 연구에서는 CART를 이용하여 음소의 지속 시간과 어절과 어절 사이의 휴지 기간을 예측하는 모델을 생성하고 이를 TTS 시스템에 적용하고자 한다.

2절에서는 실험에서 사용한 데이터와 음소 지속 시간에 영향을 미치는 변수들을 추출하기 위한 데이터의 분석에 대해서 기술하고, 3절에서는 회귀 트리를 이용하여 생성한 음소 지속 시간 예측 모델을 기술하고, 4절에서는 휴지 기간 예측 모델에 관한 기술을 한다. 그리고 마지막으로 결론을 맺는다.

2. 음성 데이터베이스와 변수의 설정

2.1 음성 데이터베이스의 구성과 음소분할

실험에 사용한 데이터베이스는 다양한 음소 문맥 정보와 문형으로 이루어진 문장을 전문 아나운서를 통해 녹음하여 자체 제작한 총 550개 문장으로 이루어진 음성코퍼스이다. 전체 음성코퍼스 중 428개 문장으로 된 음성코퍼스는 회귀 트리를 학습시키는데 사용하였고, 나머지 122개 문장으로 된 음성코퍼스는 회귀 트리 모델의 오류 정도를 실험하는 데에 사용하였다.

음성코퍼스는 auto-labeler[4]를 사용하여 변이음을 고려한 45개 음소로 분할한 뒤 수작업으로 오류를 수정하였다. 45개의 음소를 사용하여 전체 음성 코퍼스로부터 28,572개의 음소를 얻을 수 있었다. 그리고 문장과 문장 사이의 끊어 읽기 휴지 기간을 제외한 나머지 휴지기간의 개수는 836개를 얻었다.

2.2 형태소분석과 구문분석

자연스러운 운율을 생성하기 위한 변수를 설정하기 위해서는 음성코퍼스의 분석뿐만 아니라, 문장 분석과정을 통해서 나온 형태, 통사적 정보도 필요하다. 본 연구에서는 품사정보를 분석할 때, 문헌[9]에서 사용하고 있는 머리품사와 꼬리 품사를 할당하는 방식을 선택하였다. 그러나 음소단위로 품사정보를 사용하고자 할 때에는 조사나 어미의 활용 때문에 품사를 할당하는 데에 문제가 발생하므로, 머리품사와 꼬리 품사가 합쳐진 어절 품사단위를 설정하였다. 그래서 21개의 어절품사를 설정한 후 음소단위로 품사정보를 할당하였다.

2.3 특징 변수들의 설정

음소의 지속 시간과 휴지 기간을 예측하기 위해서 기존의 연구 결과를 바탕으로 다음 표1과 같이 특징 변수들을 설정하였다[5][6][7].

표 1. 특징 변수

	변 수 명	내 용
카테고리변수	Lpos Cpos Rpos	음소의 문맥 정보에 따른 좌측음소의 품사, 관측음소의 품사, 우측음소의 품사에 해당한다. 이때 음소 문맥 정보란 운율구 단위 내에서의 음소 문맥을 말한다. 관측음소가 운율구의 처음과 마지막에 해당되어 운율구 경계가 되면 경계 정보를 준다.
	Lphone Cphone Rphone	음소의 문맥 정보에 따른 좌측음소, 관측음소, 우측음소에 해당한다. 관측음소가 운율구의 처음과 마지막에 해당되어 운율구 경계가 되면 경계 정보를 준다.

카테고리변수	Ltype Ctype Rtype	음소의 문맥 정보에 따른 좌측음소의 조음양식, 관측음소의 조음양식, 우측음소의 조음양식에 해당한다.관측음소가 운율구의 처음과 마지막에 해당되어 운율구 경계가 되면 경계정보를 준다.
	BI	5단계의 Break Index 값 중에서 명백하게 끊어짐의 현상이 나타나는 4,5에 해당하는 표지만 운율구 경계로 사용하였다. [9]
	TypeSyll	관측음소가 속한 음절의 음절 유형 정보인데, 문자 표기상의 음절 유형이 아니라 발화된 후의 음절 유형에 해당한다.
	WhinEoj	어절 내에서의 관측음소의 위치정보에 해당한다. 이 값은 첫 음절, 중간 음절, 마지막 음절 세 개의 카테고리 중에 하나가 되고, 한 음절로 된 어절인 경우에는 첫 음절에 포함된다.
	WhinPhr	운율구 내에서의 관측음소의 위치정보에 해당한다. 값 설정방법은 어절 내에서의 관측음소의 위치정보에서 설정하는 방법과 동일하다.
실변수	NsylinSent	한 문장 내에서의 전체 음절 개수
	NeojinSent	한 문장 내에서의 전체 어절 개수
	NphrinSent	한 문장 내에서의 전체 운율구 개수
	NsylinEoj	한 어절 내에서의 음절 개수
	NsylinPhr	한 운율구 내에서의 음절 개수
	NphoinEoj	한 어절 내에서의 음소의 개수
	NpphoinPhr	한 운율구 내에서의 음소의 개수
	Position	한 어절 내에서의 음절의 위치

3. 음소 지속 시간의 예측

본 연구에서는 Salford Systems의 상용 CART software[8]를 사용하여 45개 음소에 대한 회귀 트리와 문장의 마지막 음절 다음의 휴지 기간 어절내의 묵음구간을 제외한 나머지 전체 휴지 기간에 대해서 회귀 트리를 생성하였다.

CART방법은 트리의 예측 오류를 최소화하는 방향으로 특징 벡터 x의 공간을 연속적으로 나누는 기법으로, 예측될 변수 y가 카테고리변수(categorical variable)인 경우는 결정 트리(decision tree)를 생성하고, 실변수(real-valued variable)일 경우에는 회귀 트리(regression tree)를 생성한다. 트리는 10-fold cross-validation 방법에 의해 생성된 트리들 중 최적 트리로 결정되고, Chou의 알고리즘에 의해 노드를 분할한다. 비단말 노드(non-terminal node)에는 노드의 문맥과 관련된 '예/아니오' 질의와 같은 제어요인이 있고, 질의에 대한 반응에 따라 두 개의 다른 가지 노드로 진행하는 과정을 반복하다가 더 이상의 분류가 어려워지면 단말 노드 (terminal node)를 만들고 이 노드에 값을 생성한다. CART 방법의 장점은 특징값으로 카테고리 변수와 실변수를 동일한 시각에서 모두 처리할 수 있다는 것과, 그 결과를 결정 트리나 회귀 트리로 만들 수 있으므로 쉽게 해석이 가능하다는 데에 있다. 예측될 변수 y에 해당하는 음소의 지속 시간과 휴지 기간은 실변수에 해당하므로 다음 절에서 설명할 특징 변수를 사용하여 회귀 트리를 생성하였다. 그런 다음 생성된 회귀 트리를 평가하기 위해서 평균제곱오류근(root mean square error(RMSE))과 예측값과 실제값 사이의 상관관계를 구하는 상관계수(correlation coefficient)를 계산해 보았다.

표 2. 각 음소트리의 RMSE

음소	RMSE(msec)	음소	RMSE(msec)
g	12.8	m	19.4
g+	17.5	n	44.1
g'	5.6	N	20.5
G	12.1	r	9.5
k	10.0	l	22.8
d	8.7	a	33.9
d+	17.7	ja	44.9
d'	7.9	v	29.7
D	11.9	jv	32.7
t	14.1	o	40.0
b	9.9	jo	47.8
b+	18.8	u	27.9
b'	12.3	ju	43.6
B	10.7	U	22.3
p	13.6	i	35.1
s	5.2	E	32.3
S	20.8	je	38.7
z	9.5	wE	42.6
z+	14.6	wa	41.0
Z	12.9	wi	38.3
c	25.7	wv	27.8
h	16.5	Wi	34.5
h+	25.1		

실험을 위해서 먼저 428 문장으로 45 개 음소 각각에 대한 회귀 트리를 학습시킨 후, 122 문장에 대해 평가를 해 보았다. 실험 문장에 대해서 실제값과 예측값의 상관계수를 구해보니 전체 상관계수는 0.84 로 의미있는 상관관계를 보였다. 표 2 는 각 음소별 트리의 RMSE 를 구한 것이다.

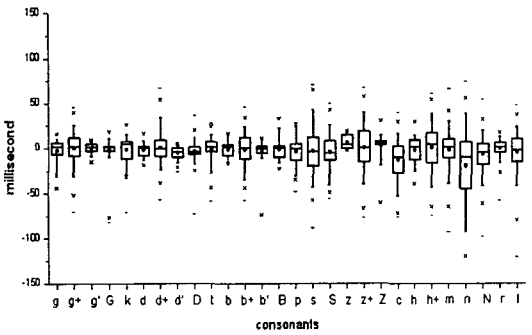

그림 1. 실제값과 예측값의 오차에 대한 v 표식 상자-수염 그림(자음)

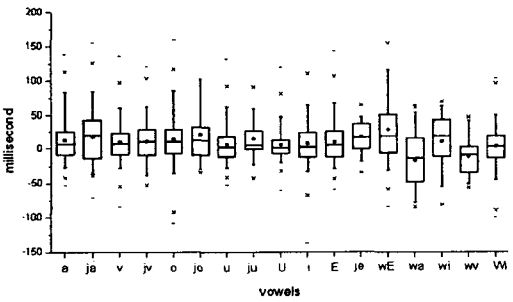

그림 2. 실제값과 예측값의 오차에 대한 v 표식 상자-수염 그림(모음)

그림 1 과 그림 2 는 각각의 트리에서 생성한 음소 지속 시간에 대한 예측값과 관측된 실제값 사이의 오차를 자음과 모음으로 나누어서 나타낸 그림이다. 음소/n/의 경우 오차가 다른 음소들에 비해 큰 데, 이것은 음소/n/이 종성과 초성에 사용될 때를 구별하지 않았기 때문에 커진 것으로 보인다. 실제 음소 지속 시간을 모델링 할 때에는 이 부분에 대한 후처리를 해주었다.

4. 휴지 기간의 예측

자연스러운 끊어 읽기를 위해서 청각적으로나 음성파형 상에서 발화의 끊김이 나타나는 부분을 표시하여 억양구를 추출하고, 억양구 사이의 휴지 기간을 측정하여 회귀 트리를 생성하였다. 그림 3 은 억양구 사이의 휴지 기간에 대한 회귀트리를 나타낸 것이다. 실험결과 실제값과 예측값 사이의 상관관계를 나타내는 상관계수는 0.63 으로 나타났다.

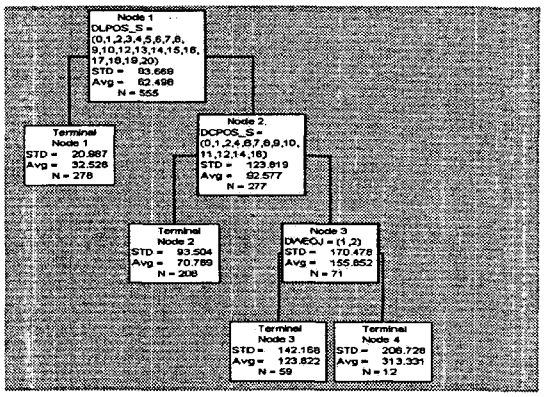

그림 3. 휴지기간에 대한 회귀트리

그림에서 먼저 root node 에서 음절의 왼쪽 품사(DLPOS_S)에 관한 질문이 나오는데, 여기에서 억양구의 경계인지를 물어보게 된다. 그런데 억양구의 경계가 아니면서 휴지 구간으로 나타나는 것이 많았다. 이는 원래 한 어절내의 음절 사이에는 휴지 기간이 없다고 가정하였으나, 조음 방법상 묵음구간이 필요한 음소가 있기 때문에 나타난 것으로 보여진다. 그래서 억양구의 첫 음절이 아니면서 묵음이 나타나는 음절에 대해서는 초성의 특성을 고려하여 묵음구간을 설정하였다. 휴지 기간은 자료의 분산도가 크기 때문에 회귀 트리 모델만으로는 정교하게 모델링되지 않아 문헌[9]에서 제안한 품사 bigram 을 사용하여 후처리를 해주었다.

5. 결론

본 연구에서는 자연스러운 합성음을 구현하기 위해서 TTS 시스템의 하위 모듈인 음소 지속 시간 예측 모델과 휴지 기간 예측 모델을 생성하였다. 각각의 상관계수는 0.84과 0.63으로 계산되었다. 휴지 기간의 모델링에 있어서는 회귀 트리 만으로는 만족할 만한 결과를 내지 못하여서, 품사 bigram을 사용하여 이를 보완하였다.

휴지 기간 모델은 음소지속 시간 모델과 깊은 관련을 가지며 자연스러운 운율을 생성하는 데 중요한 기능을 한다. 앞으로 휴지 기간 모델을 더욱 정교하게 하기 위해서 어절간의 의존관계에 관한 정보를 사용하여 휴지 기간 모델의 성능을 향상시키기 위한 노력을 계속할 예정이다.

<참고 문헌>

[1] 김형순, PC용 TEXT-TO-SPEECH 시스템 개발, 산업자원부 1차년도 중간보고서, 1998.

[2] Jan P. H. van Santen, "Deriving text-to-speech durations from natural speech," *Talking Machines : Theories, Models, and Designs*, G. Bailly, C. Benoit, and T.R. Sawallis, editors, pp.275-285, Elsevier Science, 1992.

[3] 이상호, 오영환 "CART를 이용한 운율구 추출 및 휴지 기간 모델링," *제15회 음성통신 및 신호처리 워크샵*, 15권 1호, pp.81-86, 1998.

[4] 홍성태, 김제우, 김형순 "자동 음성분할 및 레이블링 시스템의 성능향상," 대한음성학회, *말소리* 제35-36호, pp. 175-188, 1998.

[5] S.H. Lee , Y.H. Oh, "Tree-based modeling of prosodic phrasing and segmental duration for Korean TTS system," *Speech Communication*, (to appear)

[6] A.Febrer, J.Padrell, A. Bobafont "Modeling Phonetic Duration : Application to catalan TTS," *3rd International Workshop on Speech Synthesis*, Jenolan Caves, Australia, 1998

[7] B. Möbius, Jan P.H. van Santen "Modeling Segmental Durations in German Text-to-Speech Synthesis," *in Proceedings of ICSLP*, Vol. 4, pp.2395-2398, 1996.

[8] Salford System CART software [http://www.salford-systems.com]

[9] 장석복, 김형순 "Peak 파라미터와 피치 검색테이블을 이용한 억양 생성방식 연구," *제11회 한글 및 한국어 정보처리 학술발표 논문집*, 1999.