

# CART를 이용한 운율구 추출 및 휴지 기간 모델링

이 상 호<sup>○</sup>    오 영 환

한국과학기술원 전산학과

## The Modelling of Prosodic Phrasing and Pause Duration using CART

Sangho Lee<sup>○</sup>    Yung-Hwan Oh

Department of Computer Science

Korea Advanced Institute of Science and Technology

{shlee,yhoh}@bulsai.kaist.ac.kr

### 요 약

본 논문에서는 트리 기반 모델링 기법 중 하나인 CART 방법을 이용하여, 운율구 추출과 운율구 사이의 휴지 기간을 모델링하고자 한다. 모델링을 위한 특징 변수들의 유효성을 실험에 앞서 알아본 후, 생성된 트리들을 해석함으로써 제안하는 특징 변수들이 효과적임을 보인다. 음성 정보를 제외한 문서 정보만을 이용하여 실험한 결과, 운율구 경계 결정 오류율은 14.46%이었고, 휴지 기간 예측 RMSE가 132.61msec이었다.

### 1 서론

일반적으로 화자는 이야기할 때 단어들을 군집화시켜 발생하게 되고, 청자는 음의 경계 위치를 지각한 후, 화자의 의도를 해석하는데 그 정보를 이용한다[15]. 본 연구에서는 이와같이 청자로 하여금 발화의 끊김을 사각시키는 곳을 "운율구 경계", 경계의 경계 사이를 "운율구"라 부르기로 하고, 문서 음성 변환 시스템 (text-to-speech system)의 자연성을 향상시키기 위해서, 입력 문장에 대해 운율구 경계 위치를 찾고, 아울러 운율구와 운율구 사이의 휴지 기간을 예측하고자 한다.

이 두 가지 문제, 운율구 경계 위치 찾기 문제와 휴지 기간 예측 문제는 전자의 경우, 매 후보 어절이 운율구의 마지막 어절인지를 결정하는 패턴 분류 (pattern classification) 문제로 볼 수 있고, 후자의 경우, 몇 msec의 휴지 (pause)를 삽입할 것인가를 예측하는 함수 근사화 (functional approximation) 문제로 볼 수 있다. 이미 영어권에서는 전자의 문제를 해결하기 위해 비모수적 통계 모델인 CART (classification and regression trees)를[6] 이용하여 억양구 경계 (intonational phrase boundary) 위치를 높은 정확도로 찾고 있다[11, 14]. 이외에 일본어권에서는 두 종류의 휴지 시간 (긴 휴지 시간과 짧은 휴지 시간)과 구문 구조와의 관계를 보고하고 있다[12].

본 연구에서는 위의 두 가지 문제에 대해 동일한 특징 변수

들을 가지고, CART 방법을 통하여 해결하고자 한다. 우선, 품사 및 구문 정보를 포함하는 언어 정보로부터 특징 변수들을 제한한 후, 운율구 경계를 결정하는 결정 트리와 휴지 기간을 예측하는 회귀 트리를 생성한다. 실험에 앞서 특징들의 유효성에 대해 통계치를 통하여 미리 예측하고, 생성된 트리들을 해석함으로써 한국어 발화에서 운율구 경계 발생 현상에 영향을 주는 요소들을 보인다. 또한 음성 코퍼스를 배제한 문장 코퍼스만을 기반으로 하는 특징 변수들을 소개하고, 실험을 통하여 유효성을 검증한다.

본 논문의 구성은 다음과 같다. 다음 장에서는 실험에 사용된 코퍼스 및 CART 방법에 대해 간단히 알아보고, 3장에서는 운율구 추출과 휴지 기간 예측 실험 결과를 보고한다. 특히, 3.1절에서는 문장 및 음성 코퍼스를 함께 이용한 실험을 소개하고, 3.2절에서는 문장 코퍼스만을 이용한 실험을 보인다. 끝으로 4장에서 결론을 맺는다.

### 2 코퍼스 및 CART 방법

#### 2.1 코퍼스 구성 및 분석

본 연구에서 사용한 문장 및 음성 코퍼스는 다음과 같다. 여러 장르에서 발취한 400문장 (3,724어절)에 대해 기계발된 문서 분석기들[2] 통해 형태소 분석, 발음 표기 변환, 구문 분석 등을 수행하고 오류들을 수정하였다. 구문 분석은 그림 1과 같이 어절간의 관계를 지배소와 의존소로 표현하는 의존 트리로 표현한다[9]. 의존 트리는 문장내의 모든 어절들이 자신의 지배소를 하나씩 갖게 되고, 문장의 마지막 어절은 항상 자신을 지배소로 갖는다. 예를 들어, 그림 1에서 "피어"는 "꽃밭에는"과 "꽃들이"의 지배소가 되며, "있습니다"는 "피어"의 지배소가 됨을 알 수 있다. 또한, 영어의 경우와[9] 달리 한국어의 경우는 그림 1에서 보는바와 같이 지배소가 항상 의존소 뒤에 위치하는 것이 관측되며, 이를 지배소 후위의 법칙 (governor post-positioning property)이라 부른다[1]. 일반

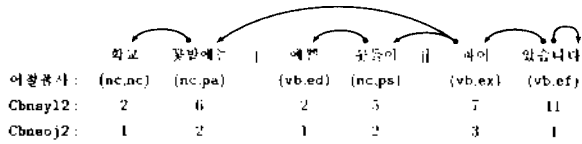


그림 1: 의존 트리의 예. 운율구 경계는 '|'로 표시되었다.

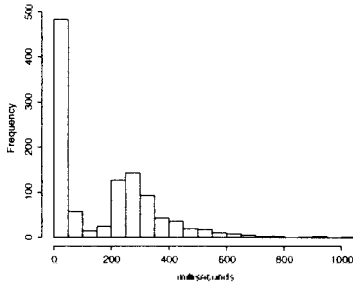


그림 2: 휴지 기간의 히스토그램.

적으로 어절의 형태소 분석 결과는 형태소 품사 열로 나타나게 되나, 본 연구에서는 어절 품사의 표현 방식을 통일화시키기 위해서 그림 1과 같이 좌품사와 우품사의 쌍으로 어절 품사를 대신하고 있으며[4], 총 23개의 품사를 사용한다\*.

전문 여성 아나운서가 방송실에서 분장 코퍼스를 발성하여, 약 33분 분량의 음성 코퍼스를 얻을 수 있었다. 음성 코퍼스에 대해 자동 음소 분할 프로그램을 이용하여 음소 경계를 표시하고[3], 분할 오류를 수정한 후 시각적으로 발화의 끊김을 지각시키는 운율구 경계를 수동 작업을 통해 얻었다. 운율구 경계 위치는 일반적으로 억양의 급변화, 마지막 음절의 장음화, 휴지의 삽입과 같은 음성 신호의 특별한 현상이 나타나게 된다. 본 연구에서는 어절 사이에만 운율구 경계가 나타난다고 가정하였고, 휴지가 삽입되지 않았을 경우에는 휴지 기간을 0msec로 가장한 후, 문장과 문장 사이의 휴지 기간을 제외 한 나머지 운율구들 사이의 휴지 기간을 측정하였다.

그림 2는 측정된 1,091개의 휴지 기간에 대한 히스토그램이다. 그림 2에서 보듯이 50msec 이하의 휴지 기간이 상대적으로 많이 관찰됨을 알 수 있다. 이는 휴지가 삽입되지 않았거나, 짧은 휴지가 삽입된 경우에도 시각적으로 운율구 경계라고 지각한다는 것을 의미하므로, 이는 “경계 앞 장음화 현상”(pre-boundary lengthening)이 휴지의 삽입보다 더욱 주요한 특징임을 간접적으로 의미한다고 볼 수 있다. 총 400문장에서 얻어진 휴지 기간들의 평균 값은 168msec이었고 표준 편차는 176msec이었으며, 전체 문장을 학습 자료 (240문장, 20분)와 실험 자료 (160문장, 13분)로 나눈 후, 두 자료의 통계치를 조사한 결과, 전체 자료의 통계치와 거의 일치하였다.

## 2.2 CART 방법

본 논문에서 구현한 CART 방법은 cost-complexity pruning 방법과 10-fold cross-validation 방법에 의해 최적 트리

\* nc:체언, vb:용언, dn:관형어, ad:부사어, it:동명어, ps:주격조사, po:목적격조사, pd:관형격조사, pv:호격조사, pa:부사격조사, pp:시술격조사, pt:수격격조사, px:보조사, pc:접속조사, ec:연결어미, ex:보조사적연결어미, ed:관형사형결어미, en:명사형결어미, ep:선어말어미, ef:어말어미, xv:동사의형접미사, xj:형용사의형접미사, xa:부사과형접미사

를 결정하고[6], Chou의 알고리즘에 의해 노드를 분할한다[8]. 노드의 불순도를 나타내는 함수로는 결정 트리일 경우, Gini 인덱스를 사용하고[6], 회귀 트리일 경우, 평균 제곱 오류를 사용한다. 최적 트리를 구할 때는 오류의 cross-validation 추정치  $R^m(T)$ 가 최소인 트리  $T$ 를 선택하는 OSE (0 standard error)법과  $R^m(T) + SE(R^m(T))$ 보다 작은 오류치를 갖는 트리를 중 트리의 크기가 최소인 트리를 선택하는 1SE 법을 모두 이용한다. CART 방법에서는 1SE 법을 제안하고 있지만[6], 최근의 비교 연구에 의하면 1SE 법이 트리를 과제거 (overpruning)한다고 알려져있다[10]. 그러나, 1SE 법으로 구한 트리의 크기가 일반적으로 더 작으므로, 본 논문에서 트리의 성능을 나타낼 때는 OSE 법으로 구한 트리의 성능을 보이고, 트리를 해석할 때는 주로 1SE 법으로 구한 트리를 이용하겠다.

본 논문에서 사용하는 표현법은 다음과 같다. 구해진 트리와 트리의 노드는 각각  $T$ 와  $t$ 로 표현하고, 트리의 단말 노드 집합은  $\hat{T}$ , 단말 노드의 개수는  $|\hat{T}|$ 로 표현한다. 학습 혹은 실험에 사용된  $N$ 개의 데이터에 대해  $R(T)$ 는 결정 트리일 경우,  $R(T) = \frac{1}{N} \sum_{t \in \hat{T}} M(t)$  ( $M(t)$ 는 노드  $t$ 에서 잘못 분류된 데이터 개수로 구해지고, 길과직으로  $N$ 개의 데이터에 대한 오류율이다. 회귀 트리의 경우에는  $R(T) = \frac{1}{N} \sum_{n=1}^N (y_n - d(x_n))^2$  ( $d(x)$ 는 입력 특성 벡터  $x$ 에 대한 트리  $T$ 의 출력값)으로 구해지고,  $\sqrt{R(T)}$ 는 평균 제곱 오류는 (root mean squared error)이 된다. 학습 데이터에 대해 cross-validation으로 구한 오류율  $\hat{R}(T)$ 는  $R^{cv}(T) \pm SE(R^{cv}(T))$ 로 표현하고, 실험 데이터에 대한 오류율은  $R^s(T)$ 로 표현한다. 한편, 회귀 트리에 대한 실험에서는 사료의 분산 정도를 얼마나 감소시켰는지 측정하기 위한 상대 평균 제곱 오류(relative mean squared error)  $RE(T) = R(T)/R(\mu)$  ( $R(\mu) = \frac{1}{N} \sum_{n=1}^N (y_n - \mu)^2, \mu = E\{y\}$ )와, 예측치와 실제치 간의 상관계수(correlation coefficient)  $r$ 도 이용된다. 이 값들도 실험 데이터에 적용한 후 얻은 값이면 각각  $RE^s(T)$ 와  $r^s(T)$ 로 표현한다.

## 3 운율구 추출 및 휴지 기간 예측

본 장에서는 운율구를 추출하고, 운율구 사이의 휴지 기간을 예측한다. 우선, 3.1절에서는 분장과 음성 코퍼스를 모두 이용하여 예측을 해보고, 3.2절에서 분장 코퍼스를만 이용하여 실험한다. 실험에 앞서, 본 연구에서 제안하는 특징들의 유효성을 알아보기 위해, 특징 변수와 예측 변수들이 모두 카테고리 변수일 경우 상호 정보 (mutual information)  $I$ 를, 모두 실변수일 경우 상관 계수 (correlation coefficient)  $r$ 을, 서로 변수의 타입이 다른 경우는 분산 분석 (ANOVA)에서 사용되는  $F$  값을 구해본다.

### 3.1 문장 및 음성 코퍼스를 이용한 예측

본 절에서는 총 18개의 특징들을 이용하여 결정 트리 및 회귀 트리를 생성한다. 사용된 특징 변수들의 상호 정보,  $F$  값, 상관 계수는 표 1과 같으며, 표에서 예측 변수가 경계 유무인 경우는 사료의 개수가 3,724개이고, 휴지 기간일 경우는 1,091개이다. 사용된 변수와 그 의미들은 다음과 같으며, 변수의 첫 글자가 D일 경우는 D 변수가 카테고리 변수임을 뜻하고 C일 경우는 실변수임을 의미한다.

제15회 음성통신 및 신호처리 워크샵(KSCSP '98 15권1호)

표 1: 특징 변수들의 상호 정보  $I$ ,  $F$  값, 상관 계수  $r$

특징 변수	경계 유무	휴지 기간
Dlpos	$I = 0.05$	$F = 19.59$
Drpos	$I = 0.30$	$F = 12.62$
Dnlpos	$I = 0.19$	$F = 19.97$
Dnrpos	$I = 0.22$	$F = 10.23$
Dgrpos	$I = 0.14$	$F = 2.68$
Dcomma	$I = 0.02$	$F = 135.22$
Deighth	$I = 0.04$	$F = 6.47$
Cneojrbeg	$F = 0.46$	$r = 0.31$
Cneojfrend	$F = 87.59$	$r = 0.24$
Cnsyltogovn	$F = 359.36$	$r = 0.37$
Cneojtogovn	$F = 343.50$	$r = 0.36$
Cnphs	$F = 34.91$	$r = 0.41$
Cnsyls	$F = 34.80$	$r = 0.41$
Cneojs	$F = 42.71$	$r = 0.40$
Cbnsyl1	$F = 508.39$	$r = 0.48$
Cbneoj1	$F = 361.55$	$r = 0.40$
Cbnsylr1	$F = 517.26$	$r = -0.10$
Cbneojr1	$F = 401.17$	$r = -0.14$

- **Dlpos, Drpos** : 관측 어절의 좌우 품사. 관측 어절의 우품사는 일반적으로 조사나 어미가 되며, 이 중 “-은”, “-는”과 같은 주격적 조사가 사용되었을 때는 운율구의 경계가 발생될 가능성이 높다.
- **Dnlpos, Dnrpos** : 관측 어절 다음에 위치하는 어절의 좌우 품사. 관측 어절 다음에 위치하는 어절의 좌우 품사는 일반적으로 체언이나 용언이 오게 되는데, 이 때 용언이 사용되었을 경우, 발화가 끊기지 않고 연결될 가능성이 높다. 만약 관측 어절이 문장의 마지막 어절일 경우 이 변수들은 NA (not-applied) 값을 갖게 되고, 400문장에서 독립어는 문장의 첫 어절로만 사용되어 Dnlpos는 독립어를 제외한 NA, 체언, 용언, 관형어, 부사어 중 하나가 된다.
- **Dgrpos** : 관측 어절의 지배소 어절 우품사. 지배소는 항상 의존소 뒤에 나타나게 되므로[1], 만약 지배소 다음에 운율구 경계가 발생되면, 의존소에서는 운율구 경계가 발생되지 않을 가능성이 높을 것이다. 다시 말해, 지배소의 우품사가 운율구 경계 발생과 밀접한 관계를 나타낸다면, 그 관계가 의존소쪽에도 영향을 미칠 것으로 생각할 수 있다. 400문장에서 관형어, 독립어 혹은 호격 조사가 우품사인 어절들이 의존소를 갖은 경우가 없었으므로, Dgrpos는 이들을 제외한 14개 품사 중 하나가 된다.
- **Dcomma** : 쉼표의 유무. 일반적으로 어절 다음에 쉼표가 있으면, 운율구 경계일 가능성이 높게 된다[13].
- **Deighth** : 문장내에서 관측 어절의 상대적 위치. 기존의 연구에 의하면, 운율구 경계는 비슷한 크기의 단위로 발화를 나누는 것으로 알려져 있다[13]. 관측 어절의 상대적 위치를 1에서 8사이의 정수 값으로 양자화하여 사용하고, 주의할 점은 이 변수가 실변수가 아닌 카테고리 변수라는 점이다.

표 2: 운율구 추출-I (성능)

학습 ( $N=2286$ )	실험 ( $N=1438$ )
$\hat{R}(T) = 0.1351 \pm 0.0071$	$R^s(T) = 0.1481$
$ \hat{T}  = 11$	

표 3: 운율구 추출-I (confusion 행렬)

	경계	부경계	백분율
경계	455	135	77.11%
부경계	78	770	90.80%

- **Cneojrbeg, Cneojfrend** : 문장의 첫 어절과 마지막 어절부터 몇 번째 어절에 해당하는지에 대한 정보. 관측 어절이 문장의 앞 부분에 가깝거나 혹은 끝 부분에 가까우면, 운율구 경계가 발생되지 않을 가능성이 높다[14].
- **Cnsyltogovn, Cneojtogovn** : 시배소까지의 음절과 어절 개수. 관측 어절  $w_i$ 와 지배소  $w_j$  ( $i < j$ )까지의 음절 혹은 어절 개수가 많다면, 어절들  $w_{i+1}, \dots, w_j$ 의 중심 크기가 커진다는 의미이므로 관측 어절  $w_i$  다음에 운율구 경계가 발생될 가능성이 높다. 또한, 이미 운율구 경계라면 휴지기간이 길어질 가능성이 높다.
- **Cnphs, Cnsyls, Cneojs** : 문장내의 음소, 음절, 어절 개수. 문장을 이루는 음소, 음절 혹은 어절 개수가 작으면 그 문장의 발화 기간도 짧을 것이며, 결과적으로 그 문장내에 운율구 경계의 발생 빈도도 역시 작을 것이다. 이 특징 값들은 문장내의 모든 어절에서 동일한 값이 사용된다.
- **Cbnsyl1, Cbneoj1** : 이전 운율구 경계로부터 관측 어절까지의 음절 개수와 어절 개수. 이전 운율구 경계로부터 관측 어절까지의 음절 혹은 어절 개수가 많을수록 운율구 경계가 발생될 가능성이 높다. 이는 일반적으로 사람이 발화를 할 때 이전 운율구 경계로부터 일정 시간이 지나면 다시 운율구 경계를 발생시킨다는 당연한 사실에 바탕으로 한 특징 값이다. 그리고, 이 값은 표 1에 나타나듯이 휴지 기간과의 상관관계가 0.48로 다른 실 변수들에 비해 휴지 기간과의 상관도가 가장 높다. 여기서 사용된 이전 운율구 경계는 음성 코퍼스에 표시한 운율구 경계를 사용한다.
- **Cbnsylr1, Cbneojr1** : Cbnsyl1, Cbneoj1을 전체 문장의 음절과 어절 개수로 나눈 값. 일반적으로 영어권에서는 이전 운율구의 단어 개수로 나눈 값을 특징 값으로 이용한다[13]. 이 때 문장의 첫부분에서는 이전 운율구가 존재하지 않아 그 값을 구할 수 없게 되고, 이렇게 특징 변수들 중 관측되지 못한 결어 변수가 있을 때 CART에서는 대리 변수를 이용하여 이를 해결한다[6]. 본 논문에서 구현한 프로그램은 결어 변수를 해결하지 못하므로 문장 길이를 대신 이용한다. 특별히 문장 길이를 사용한 이유는 위의 Deighth 설명에서 보듯이 운율구의 길이가 전체 문장의 길이와 선형적인 관계에 있다는 가정을 바탕으로 한다.

초청논문: CART를 이용한 운율구 추출 및 휴지기간 모델링

표 4: 휴지 기간 예측-I

학습 (N=661)	실험 (N=430)
$\hat{R}(T)$ -- 16850.06 ± 1512.46	$R^{ts}(T)$ -- 16716.94
$ \hat{J} $ = 10	$\sqrt{R^{ts}(T)}$ = 129.29
	$RE^{ts}(T)$ = 0.58
	$r^{ts}(T)$ = 0.65

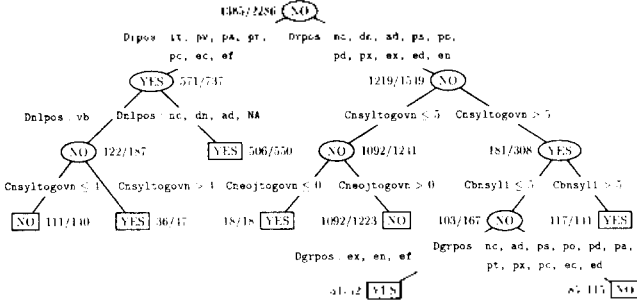


그림 3: 운율구 경계 결정 트리-I (1SE 법).

위의 특징들을 이용하고, 240문장 (2,286어절)으로 운율구 추출 결정 트리를 학습시키고 160문장 (1,438어절)으로 실험한 결과, 표 2와 같이 약 85%의 정확률을 얻었다. 결정 트리의 confusion 행렬은 표 3과 같았으며, 경계 유무 결정 백분율의 평균값은 83.96%이었다. 운율구 사이의 휴지 기간 예측 실험에서는 표 4와 같은 결과를 얻었고, 표에서 보는 바와 같이 상대 평균 제곱 오류값이 0.58, 실제 휴지기간과 예측 휴지기간과의 상관값이 0.65로 자료의 분산도가 큰 휴지 기간 값도 효과적으로 예측할 수 있었다.

학습된 결정 트리과 회귀 트리는 그림 3, 4와 같았으며, 각 트리들은 1SE 법에 의해 구해진 트리들이다. 결정 트리의 노드 좌우에 있는  $x/y$ 에서  $y$ 는 루트 노드로부터 그 노드에 도달한 자료의 개수이고,  $x$ 는 노드의 레이블 (YES 혹은 NO)에 해당하는 자료의 개수이다. 결정 트리에서 사용된 특징 변수들을 관찰해 보면, 카테고리 변수들 중 가장 상호 정보 값이 큰  $Dgrpos$ ,  $F$  값이 다른 실변수들보다 큰  $Cnsyltogovn$ 와  $Cbsnyll$ 이 주로 사용된 것을 알 수 있다. 루트 노드의 질문은 관측 어절의 우품사가 독립어 (it), 호격 조사 (pv)와 같이 변화의 끊임이 당연히 발생할 것이라고 예상되는 품사들과 체언 (nc), 관형사형 전성어미 (ed)와 같이 그 다음 어절과 연속될 가능성이 높은 품사들로 크게 나뉘어졌음을 알 수 있다. 그리고,  $Cnsyltogovn$ 와  $Cbsnyll$ 이 사용된 노드에서는 예상했던 바와 같이 모두 그 값들이 클 때 운율구 경계 발생률이 높았다. 한편,  $Cneojtogovn$ 가 사용된 노드에서  $Cneojtogovn$ 가 0일 경우 운율구 경계가 발생되었는데, 이는 문장의 마지막 어절이 항상 자신을 지배소로 갖는다는 가정에 기인한 것이다.

관측 어절 다음에 위치하는 어절의 좌품사 ( $Dnlpos$ )가 사용된 노드를 살펴보면, 용언 (vb)이 사용되었을 때, 운율구 경계가 발생될지 않을 가능성이 높았다. 이러한 이유는 다음과 같이 해석할 수 있다. 우선, 트리에서 보듯이 지배소까지의 음절 개수가 많을수록 운율구 경계의 발생 빈도가 높았으며, 이는 지배소까지의 어절 개수의 경우도 동일한 특성을 보인다. 표 1에서  $Cnsyltogovn$ 와  $Cneojtogovn$ 의  $F$  값을 살펴보면, 각각 359.36, 343.50으로 높은 값을 갖고 있다. 실제로 코퍼스를 조사해본 결과,  $Cneojtogovn$  값이 커짐에 따라 운

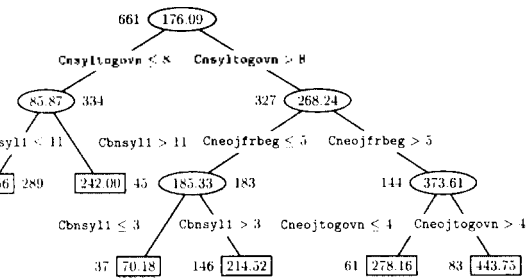


그림 4: 휴지 기간 회귀 트리-I (1SE 법).

율구 발생 확률이 높아졌다.  $Cneojtogovn$  값과 발생 확률과의 관계를 더욱 자세히 살펴보면, 관측 어절이 다음 어절의 의존소가 되는 경우 ( $Cneojtogovn$  값이 1인 경우)의 빈도가 총 400문장에서 2195번 나타났으며, 그 중 운율구 경계가 발생된 경우가 279번으로, 발생 확률은 0.127로 가장 낮은 확률을 나타내었다. 한편,  $Dnlpos$ 와  $Cneojtogovn$ 와의 관계를 살펴보면,  $Dnlpos$ 가 용언 (vb)인 경우가 총 1191번 있었으며, 그 중 관측 어절이 바로 다음 어절의 의존소가 되는 경우가 1059번으로, 확률  $P(Cneojtogovn = 1|Dnlpos = vb)$ 는  $1059/1191 \approx 0.89$ 가 된다.  $Dnlpos$ 가 체언 (nc)의 경우,  $1109/1859 \approx 0.59$ , 관형어 (dn)의 경우,  $1/88 \approx 0.01$ , 부사어 (ad)의 경우,  $26/186 \approx 0.13$ 이이었으므로, 결국 용언의 경우가 가장 높은 확률로 관측 어절이 다음 어절의 의존소가 되며, 아울러 운율구 경계 발생 가능성도 낮아지게 된다.

지배소 우품사 ( $Dgrpos$ )가 사용된 노드를 살펴보면, 보조적 연결어미 (ex), 명사형 전성어미 (en), 어받어미 (ef)가 지배소의 우품사일 때 운율구 경계가 발생할 가능성이 높았다. 이러한 현상의 이유를 트리의 루트 노드로부터  $Dgrpos$ 가 사용된 노드까지의 경로를 해석함으로써 알아보자. 우선 관측 어절 우품사가 운율구 경계를 발생시킬 가능성이 낮으며 (루트 노드 질문), 지배소까지의 거리가 커서 ( $Cnsyltogovn > 5$ ) 운율구 경계가 발생할 가능성이 높아진다. 그러나, 이전 운율구 경계로부터 일정 시간이 지나지 않았기 때문에 ( $Cbsnyll \leq 5$ ) 관측 어절에서 운율구 경계를 발생시킬 것인지를 지배소 우품사를 조사함으로써 결정한다. 즉, 지배소 우품사가 보조적 연결어미, 명사형 전성어미와 같이 지배소 다음에 운율구 경계가 발생될지 않을 가능성을 높이는 품사가 사용되었을 경우에는, 비록 어절 운율구 경계로부터 일정 시간이 지나지 않아도, 지배소 이후 어절 쪽에서 운율구 경계가 발생할 것으로 가정하고 (이는 화자로 하여금  $Cnsyltogovn$ 가 나타내는 어절의 문장 크기보다 더 큰 문장이 될 가능성이 높다고 판단하게 한다.) 관측 어절에서 운율구 경계를 발생시키게 된다. 실제로 100문장에서 보조적 연결어미와 명사형 전성어미는 각각 149번, 22번 나타났고, 이 중 각 어미 다음에 운율구 경계가 발생된 횟수는 각각 10번, 1번으로, 총 17개의 우품사 중 0.236의 확률을 갖은 관형어를 제외하고 가장 운율구 경계 발생 확률이 낮았다. 특징 변수를 소개할 때 설명하였듯이 관형어는  $Dgrpos$ 로 사용되지 않았으므로, 실제로는 운율구 경계 발생 확률이 명사형 전성어미의 경우 0.045로 가장 확률이 낮았으며, 보조적 연결어미는 0.067로 두번째로 확률이 낮았다. 이상에서 지배소 우품사가 관측 어절의 운율구 발생과 관계가 있음을 알 수 있다. 결론으로 0SE 법에 의한 결정 트리에서는 이외에  $Cbneojr1$ 을 포함한 총 7개의 변수가 이용되었다.

그림 4의 회귀트리에서 사용된 특징 변수들을 살펴보면 상관 계수가 상대적으로 큰  $Cnsyltogovn$ 와  $Cbsnyll$ 이 주로

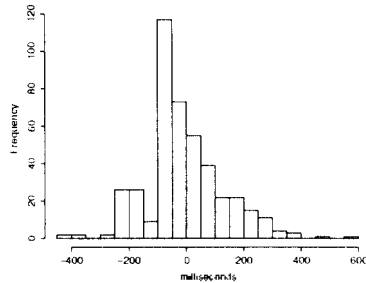


그림 5: 휴지 기간 예측 오류의 히스토그램.

사용되었음을 알 수 있다. 또한 트리에서 Cnsyltogovn와 Cbsnyl1의 값이 클수록 휴지 기간이 길어졌고, 이는 특징 변수를 소개할 때 설명하였듯이 이질 군집의 크기와 휴지 기간이 양의 상관 관계를 갖는다는 것을 나타낸다. 한편, 트리에서 Cneojr2beg이 사용된 노드를 살펴보면, 이 값이 클수록 휴지 기간이 길어지는 것을 관측할 수 있는데, 이는 다음과 같이 해석할 수 있다. 기존의 연구에 의하면, 말화에 사용된 모든 음소들의 지속 시간 합과 모든 휴지 기간의 합은 선형적으로 비례한다고 알려져 있다[5]. 그러므로, 위의 특징들 중 분장을 이루는 음소 개수 (Cnphs)는 음소들의 지속 시간 합과 양의 상관 관계가 있을 것이며, 이는 비록 휴지 기간 값이 아니더라도 각 휴지 기간과 양의 상관 관계를 갖는다고 가정할 수 있어서, 표 1에 상관 계수가 0.41인 것을 볼 수 있다. 만약 문장 처음부터 관측 어절까지의 거리인 Cneojr2beg의 값이 크다면, 이는 문장이 길다는 것을 의미하며, 결과적으로 음소의 개수가 많아서 휴지 기간도 길어지게 된다. 실제로 Cneojr2beg와 Cnphs의 상관 계수를 구한 결과, 0.65의 양의 상관도를 보였다. OSE 법에 의한 회귀 트리는 이외에 변수 Cbneoj1과 Dcomma를 포함한 총 6개의 변수가 사용되었으며, 물론 유효성이 있을 때 휴지 기간이 길어졌다.

그림 5는 실제 휴지 기간과 회귀 트리를 이용하여 예측한 휴지 기간과의 차이에 대한 히스토그램이다. 그림에서 오류값이 0msec가 아닌 -50msec 정도가 가장 많이 발생되었는데, 이는 2.1절에서 설명하였듯이 휴지가 삽입되지 않았을 경우에 휴지 기간을 0msec로 가정하였으나, OSE 법에 의한 회귀 트리가 예측하는 가장 짧은 휴지 기간은 50.97msec로 차이를 보이고 있기 때문이다.

### 3.2 문장 코퍼스만을 이용한 예측

앞 절에서 사용한 특징 변수들 Cbsnyl1, Cbneoj1, Cbsnyl1r1, Cbneojr1은 음성 코퍼스로부터 얻은 정보들이다. 본 절에서는 이 네 가지 특징들을 대신할 수 있는 변수들을 구분 정보를 기반으로 다음과 같이 제안한다. 이들의 F 값과 상관 계수를 표 5에 보인다.

- Cbsnyl2, Cbneoj2: 관측 어절  $w_i$ 와 그 앞에 위치하는 임의의 어절  $w_k$  ( $k < i$ )에 대하여, 그 사이에 위치하는 모든 어절들  $w_l$  ( $k < l < i$ )이 항상 어절  $w_{i-1}$ 의 지배 소임을 만족할 때, 가장 앞부분에 위치하는 어절  $w_k$ 와 어절  $w_i$  사이의 음절 개수와 어절 개수. 이 특징 값은 관측 어절까지의 국부적인 어절의 군집 크기를 나타낸다. 이 값이 크면 관측 어절까지 말화의 끊김이 없을 가능성이 높게 되어 결과적으로 관측 어절에서는 율율구 경계

표 5: 대응되는 특징 변수들의 F 값, 상관 계수 r

특징 변수	경계 율율	휴지 기간
Cbsnyl2	$F = 544.38$	$r = 0.44$
Cbneoj2	$F = 501.83$	$r = 0.39$
Cbsnyl2r2	$F = 528.77$	$r = -0.11$
Cbneoj2r2	$F = 469.39$	$r = -0.15$
Cbsnyl3	$F = 245.26$	$r = 0.36$
Cbneoj3	$F = 222.49$	$r = 0.34$
Cbsnyl3r3	$F = 472.69$	$r = -0.07$
Cbneoj3r3	$F = 436.04$	$r = -0.11$

표 6: 율율구 추출-II (성능)

학습 (N=2286)	실험 (N=1438)
$\hat{R}(T) = 0.1364 + 0.0071$	$R^2(T) = 0.1446$
$ \hat{T}  = 15$	

가 발생할 가능성이 높다. 이해를 돕기 위해 그림 1에 각 어절의 Cbsnyl2와 Cbneoj2를 보인다.

- Cbsnyl2, Cbneoj2: Cbsnyl2, Cbneoj2를 전체 문장의 율율과 어절 개수로 나눈 값.
- Cbsnyl3, Cbneoj3: 관측 어절  $w_i$ 가 지배하는 어절들  $w_m$  ( $m < i$ )중 가장 앞 부분에 위치하는 어절  $w_m$ 까지의 음절 및 어절 개수. 이 특징 값은 Cnsyltogovn와 Cneojtogovn를 지배소 관점에서 본 것으로, 이 값이 크면 관측 어절에서 율율구 경계가 발생할 가능성이 높다.
- Cbsnyl3r3, Cbneojr3: Cbsnyl3, Cbneoj3를 전체 문장의 율율과 어절 개수로 나눈 값.

앞 절에서 사용한 네개의 특징 변수 대신에 위의 특징 변수들을 이용하여 다시 트리들을 학습시킨 결과, 율율구 추출 성능, confusion 행렬, 휴지 기간 예측 성능이 각각 표 6, 7, 8과 같았다. 경계 율율 결정 백분율의 평균값은 84.48%였고 휴지 기간 예측 실험에서 상관값이 0.63으로, 표 2, 3, 4와 비교할 때, 트리의 성능이 거의 비슷함을 알 수 있다. 트리의 크기는 결정 트리의 경우 단말 노드 개수가 11개에서 15개로 증가했으며, 회귀 트리의 경우는 오히려 10개에서 6개로 감소하였다. 특히, 본 절에서 OSE 법으로 구한 회귀 트리가 앞 절에서 ISE 법으로 구한 회귀 트리와 트리의 크기가 동일하였다.

그림 6, 7은 제안한 특징 변수들을 이용한 결정 트리과 회귀 트리이다. 그림 6과 그림 3을 비교하면, 앞 절에서 사용된 변수 Cbsnyl1 대신에 Cbsnyl2가 사용되었음을 알 수 있다. 또한 그림 7과 그림 4를 비교하면, Cbsnyl1 대신에 Cbsnyl2가 사용되었음을 알 수 있다. 만약 Cbsnyl2와 Cbsnyl1간의 율율상 차이가 있다면, 결정 트리과 회귀 트리의 형태가 앞 절의 트리들과 많은 차이를 보였을 것이다. 그러나, 그림에서 보듯이 앞 절의 트리들에서 Cbsnyl1이 사용된 노드만이 교체된 것으로 이루어, 본 절에서 제안한 8개 특징들의 율율성이 앞 절의 음성 코퍼스에 기반을 둔 특징들의 율율성과 유사하다고 생각할 수 있다. 참고로, Cbsnyl1과 Cbsnyl2와의 상관 계수를 구한 결과, 0.61을 얻을 수 있었다.

## 초청논문: CART를 이용한 운율구 추출 및 휴지기간 모델링

표 7: 운율구 추출-II (confusion 행렬)

	경계	무경계	백분율
경계	464	126	78.64%
무경계	82	766	90.33%

표 8: 휴지 기간 예측-II

학습 (N=661)	실험 (N=130)
$\hat{R}(T) = 17951.78 \pm 1507.38$	$R^2(T) = 17586.95$
$ \hat{T}  = 6$	$\sqrt{R^2(T)} = 132.61$
	$RE^2(T) = 0.61$
	$r^2(T) = 0.63$

### 4 결론

본 논문에서는 CART 방법을 이용하여 운율구 추출 및 휴지 기간을 모델링하였다. 실험을 통하여 제안한 특징 변수와 이를 이용한 CART 방법이 운율 모델링에 효과적인임을 알 수 있었다. 그러나, 운율구 추출 정확률이 약 85% 정도로 아직 자연스러운 운율을 기대하기에는 좀 더 높은 정확률이 요구된다. 최근에 CART의 정확률을 높일 수 있는 방법들이 제안되고 있으며[7], 앞으로 CART 방법의 향상을 통하여 정확률을 높일 계획으로 있다.

### 참고 문헌

- [1] 김 창현, 김 재훈, 서 정연, "지배가능경로를 이용한 오른쪽 우선 구문분석," 제 5회 한글 및 한국어 정보처리 학술 발표 논문집, pp. 35-44, 1993.
- [2] 이 상호, 오 영환, 서 정연, "한국어 문서 음성 변환 시스템을 위한 문서 분석기", 한국음향학회지, Vol. 15, No. 3, pp. 50-59, 1996.
- [3] 이 상호, 오 영환, "음성 코덱으로부터 TTS 시스템을 위한 단위 음성 자동 생성", 한국음향학회 학술발표대회 논문집 제 16권 1호, pp. 293-296, 1997.
- [4] 이 상호, 오 영환, "CART를 이용한 운율구 추출 및 음소 지속 시간 모델링", 한국음향학회 학술발표대회 논문집 제 17권 1호, pp. 135-138, 1998.
- [5] G. Bailly, "Integration of Rhythmic and Syntactic Constraints in a Model of Generation of French Prosody," *Speech Communication*, Vol. 8, pp. 137-146, 1989.
- [6] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, *Classification and Regression Trees*, Wadsworth Statistics/Probability Series, Belmont, CA, 1981.
- [7] L. Breiman, "Bagging Predictors," *Machine Learning*, Vol. 24, pp. 123-140, 1996.

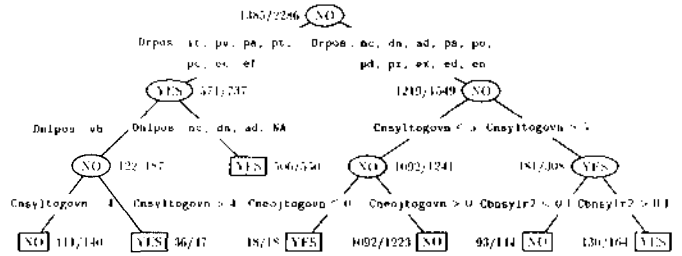


그림 6: 운율구 경계 결정 트리-II (1SE 법).

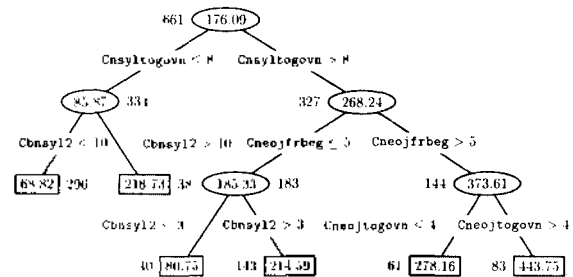


그림 7: 휴지 기간 회귀 트리-II (0SE 법).

- [8] P.A. Chou, "Optimal Partitioning for Classification and Regression Trees," *IEEE Trans. on PAMI*, Vol. 13, No. 4, pp. 304-354, 1994.
- [9] M.A. Covington, A dependency parser for variable-word-order languages, *Research Report AI-1990-01*, Artificial Intelligence Programs, The University of Georgia, 1990
- [10] F. Esposito, D. Malerba and G. Semeraro, "A Comparative Analysis of Methods for Pruning Decision Trees," *IEEE Trans. on PAMI*, Vol. 19, No. 5, pp. 476-491, 1997.
- [11] J. Hirschberg and P. Prieto, "Training intonational phrasing rules automatically for English and Spanish text-to-speech," *Speech Communication*, Vol. 18, pp. 281-290, 1996.
- [12] N. Kaiki and Y. Sagisaka, "Pause Characteristics and Local Phrase-Dependency Structure in Japanese," *Proceedings ICSLP*, pp. 357-360, 1992.
- [13] M. Ostendorf and N. Veilleux, "A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location," *Computational Linguistics*, Vol. 20, No. 1, pp. 27-54, 1994.
- [14] M.Q. Wang and J. Hirschberg, "Automatic Classification of Intonational Phrase Boundaries," *Computer Speech and Language*, Vol. 6, pp. 175-196, 1992.
- [15] C.W. Wightman, S. Shattuch-Hufnagel, M. Ostendorf and P.J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Am.*, Vol. 91, No. 3, pp. 1707-1717, 1992.