음성합성을 위한 C-ToBI기반의 중국어 운율 경계와 F0 Contour 생성

김승원(포항공대), 정옥(포항공대),
이근배(포항공대), 김병창(대구가톨릭대)

<차 례>

<Abstract>

Chinese Prosody Generation Based on C-ToBI Representation for Text-to-Speech

Seungwon Kim, Yu Zheng, Gary Geunbae Lee, Byeongchang Kim

Prosody modeling is critical in developing text-to-speech (TTS) systems where speech synthesis is used to automatically generate natural speech. In this paper, we present a prosody generation architecture based on Chinese Tone and Break Index (C-ToBI) representation. ToBI is a multi-tier representation system based on linguistic knowledge to transcribe events in an utterance. The TTS system which adopts ToBI as an intermediate representation is known to exhibit higher flexibility, modularity and domain/task portability compared with the direct prosody generation TTS systems. However, the cost of corpus preparation is very expensive for practical-level performance because the ToBI labeled corpus has been manually constructed by many prosody experts and normally requires a large amount of data for accurate statistical prosody modeling. This paper proposes a new method which transcribes the C-ToBI labels automatically in Chinese speech. We model Chinese prosody generation as a classification problem and apply conditional Maximum Entropy (ME) classification to this problem. We empirically verify the usefulness of various natural language and phonology features to make well-integrated features for ME framework.

* Keywords : Prosody, Text-to-Speech, ToBI, Maximum Entropy (ME)

1. 서 론

TTS(text-to-speech) 시스템에서 중요한 문제 중 하나는 운율구(prosodic phrase), 피치(F0, pitch) 곡선, 분절 기간 패턴을 알맞게 다루는 것이다. 특히 TTS 시스템에서 운율구로 나누는 것과 피치곡선을 생성하는 것은 자연스런 음성을 만드는데 가장 중요하다. 언어를 연구하는 사람들은 음성 언어는 음운 구, 억양 구, 발성을 포함하는 운율 요소들의 계층 구조로 되어있다고 말하며[1], 문자 언어는 운율 요소들과는 다른 단어들이나 구 같은 문법적 요소들로 이루어진다고 말한다. 그러나 우리는 문법적 정보가 운율구를 추정하기 위한 중요한 단서를 제공할 것이라고 가정한다. 운율 경계(phrase break) 추정을 위해 Recurrent Neural Network[2], Hidden Markov Model[3] 같은 많은 기술들이 소개되었고, Zhao와 Tao[4]는 두 개의 전형적인 규칙 학습 알고리즘(C4.5 와 TBL)으로 자동 규칙 학습 접근을 제안했다. 이들은 품사 자질(POS feature), 어휘(lexical) 자질, 길이 자질을 사용하여 87.9% 이상의 정확도를 보였다. 또한 묶음(chunking) 자질을 추가하여 90% 이상의 성과를 보이기도 했다. 그러나 그들은 정확도를 평가하기 위해 단지 두 종류의 운율적 구조를 사용했다.

중국어는 음조의 언어(tonal language)이기 때문에 보통 음절(syllable)이 처리를 위한 기본 운율 요소로 배정되며, 각 음절은 톤과 비교적 안정된 피치 곡선을 갖는다. 그러나 피치 곡선은 자연스런 음성 안에서 문맥적 정보에 의해 고립된 음절이 다른 음절에 영향을 주면서 변형된다. 그러므로 중국어에서 피치 악센트 추정은 어려운 문제이다. HMM, neural network, decision tree, bagging, boosting 같은 많은 기계 학습 기술들이 피치 악센트 추정을 위해 소개되었었다. Xuejing Sun[5]은 4 종류의 영어 피치 악센트 레이블로 전체 결정 트리 접근법을 제안했다. 그들의 방법은 기준 정확도 보다 12.28%의 성능이 향상된 80.50%의 정확도를 보이지만 텍스트 자질만 사용했다. Michell L. Gregory와 Yasemin Altun[6]는 두 종류 영어 피치 악센트 레이블로 CRF 기반의 접근법을 제안했고, 기준 정확도 보다16.86%의 성능 향상을 가져왔다.

C-ToBI는 매개 표현으로 성능의 저하 없이 시스템 레벨의 모듈성, 호환성, 이식성을 증가시킨다. 그러나 음성 코퍼스에서 C-ToBI 레이블링은 많은 노력과 시간이 들어간다. 그러므로 우리는 자동 C-ToBI 레이블링 방법을 제안하여 이를 해결하고자 한다. 우리는 운율 경계 추정과 피치 곡선 추정을 다루고 조건부 ME (maximum entropy) 모델기반을 C-ToBI 레이블에 적용했으며, 다양한 종류의 언어적, 음성적 정보들을 자질(feature)의 형태로 표시했다.

이 논문은 다음과 같이 구성된다. 2장에서는 운율 시스템에 대한 기존의 연구에 대해 기술하고, 3장에서는 자동 C-ToBI 레이블링 시스템에 대해, 4장은 우리의 C-ToBI기반의 운율 경계 추정, 피치 악센트 추정, 피치 곡선 생성 방법에 대해 소

개한다. 우리가 제안한 방법의 효과는 5장에서 실험 결과로 보여 지며 마지막으로 6장에서 결론을 내린다.

2. C-ToBI 시스템

톤과 경계 색인 시스템(Tones and Break Indices system (ToBI))은 억양의 패턴과 영어 발성의 운율적인 측면[7]을 전사하기 위한 시스템이다. C-ToBI[1]는 표준(Mainland) 중국어의 운율 전사 규약이다. ToBI 시스템은 다중 층 (tier)로 구성되는데 발성의 운율적 사건을 각 레이블로 표현한다. C-ToBI 레이블링 시스템은 영어 ToBI 시스템과 원리적인 측면에서 비슷하다. 우리는 음성 합성과 음성 인식을 위해 담화와 자연스런 대화의 운율적 정보를 제공할 목적으로 운율 레이블링에 대한 C-ToBI 규약을 <표 1>에서 정하였다.

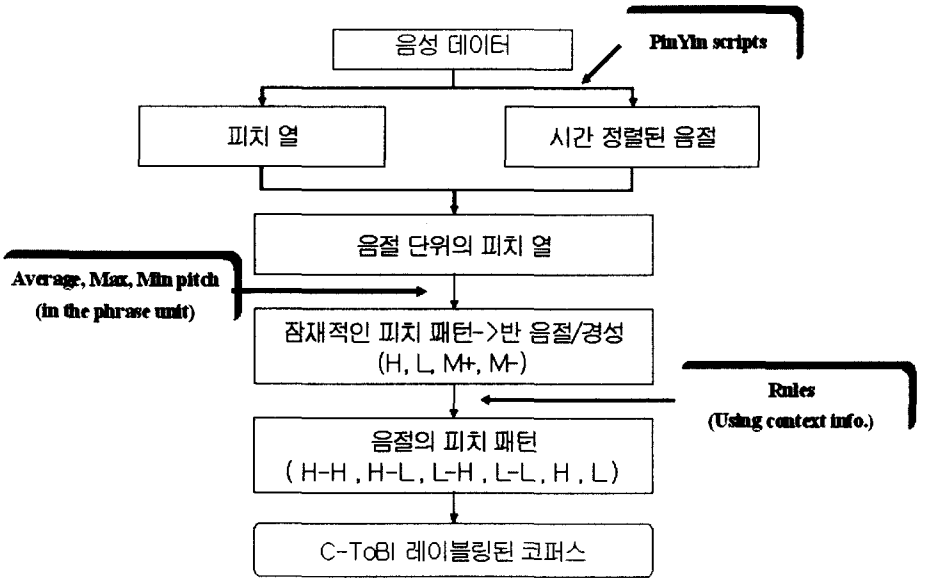
<표 1> C-ToBI (버전2.0)

레이블	C-ToBI에 관한 설명
$	침묵
@	축소
&	과도기의 음조(transitional tone)
S, Q, I, E	S: 진술; Q: 의문; I: 명령; E: 감탄;
Stress index: 0-4	5개의 종류로 강세 (stress tier)수준을 표기한다. 0-4: non-stress, PW, MIP, MAP, IU.
Break index: 0-4	5개의 종류로 운율 경계 (break)수준을 표기한다. 0-4: non-break, PW(prosodic word), MIP(minor phrase), MAP(major phrase), IU(prosodic group).
H-L, L-H, H-H, L-L	음조와 억양의 자질(tonal feature)들
H- L-	경성 음조(neutral tone)와 억양의 자질들
H%, L% (T&I)	경계 음조(boundary tone)
^, ^^	^ 증진; ^^ 폭넓은 증진
!, !!	! 감축; !! 폭넓은 감축
R	Register shifting

1) http://www.cass.net.cn/chinese/s18_yys/yuyin/english/ctobi/ctobi.htm

3. 자동 C-ToBI 레이블링

C-ToBI 톤은 마치 피치 곡선처럼 다양한 언어적 지식과 억양의 변화를 담고 있기 때문에 TTS 시스템에서 중요한 부분이다. 하지만 C-ToBI 레이블링 코퍼스를 구축하려면 음성학 지식을 갖춘 전문가들이 음성 코퍼스를 들어보며 레이블링을 해야 하기 때문에 매우 어렵고 시간이 많이 든다. 게다가 수동적으로 코퍼스를 구축하면 사람에 따라 다르기 표기 될 수 있기 때문에 일관성이 결여 되고, 통계학적 방법을 사용해 C-ToBI 톤을 레이블링하려면 많은 코퍼스를 필요로 한다. 그렇기 때문에 우리는 이에 대한 대안으로 자동적으로 C-ToBI를 레이블링하는 방법[8]을 제안한다. <그림 1>은 자동 C-ToBI 톤 레이블링 과정을 보여준다.



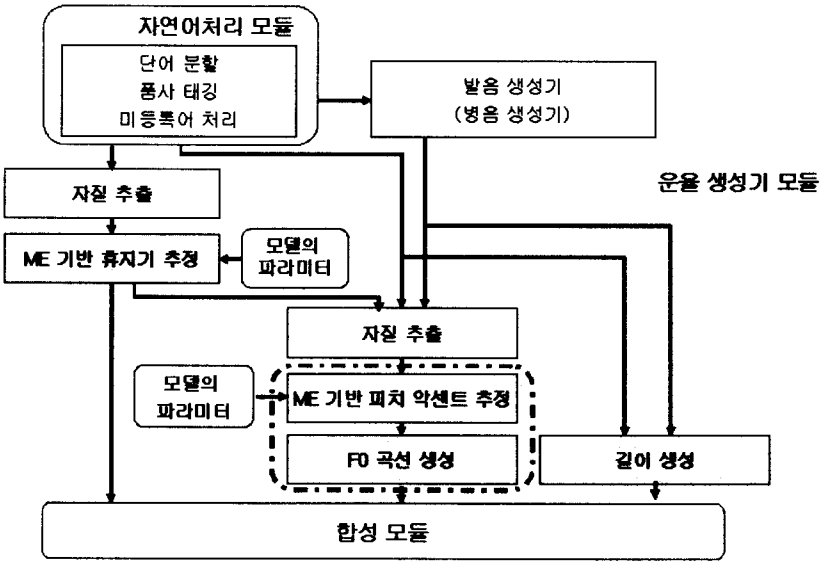
<그림 1> 자동 C-ToBI 톤 레이블링

우선 음성 데이터로부터 피치를 추출한 뒤에 병음 단위(syllable script)로 음성을 정렬하고 추출된 피치 단위로 병음을 정렬한다. 병음 길이에 따라 피치 패턴이 중성(neutral tone)인지 두 개의 반 음절(half-syllable)로 이루어진 톤인지 구별하고, 각 운율구(prosodic phrase) 내의 피치 열의 최대, 최소, 평균값을 계산한다. 이 값은 중성 톤과 반 음절이 L 톤인지 H 톤인지 결정하는 기준이 된다. C-ToBI 레이블링 과정에서는 사람이 억양의 변화를 인지하는 한계가 있기 때문에 자연적으로 애매한 구역이 나타나는데 이 애매한 구역은 H 톤에 가까운 M+ 톤과 L 톤에 가까운 M- 톤으로 적은 뒤 문맥적 정보를 이용해 다시 H 나 L 톤으로 적는다. 그 다음 각각의 반 음절 쌍을 하나의 피치 패턴 형태로 합친다. 이 같은 계산 과정은 음성

데이터에서 툴을 사용하여 자동으로 피치 열을 추출한 뒤 HTK툴을 사용하여 시간 정렬이 된 음절을 추출하고 이를 다시 음절 단위의 피치 열로 변환함으로써 구할 수 있다. 그 다음 각 음절 단위로 피치의 최고치, 최소치, 평균치를 계산하여 그것을 기준으로 L 또는 H 레이블을 할당한다. 그러면 나온 결과로 위의 과정을 거치면서 C-ToBI 레이블링된 코퍼스를 얻을 수 있다.

4. C-ToBI기반 운율 추정 모델

<그림 2>는 운율 추정 모델의 전반적 구조다. 이 모델은 휴지기(운율 경계) 추정, 피치 악센트 추정, F0(피치) 곡선을 생성하는 부분으로 나눌 수 있다. 본 연구에서는 기존에 개발된 단어 분할과 품사 태깅 시스템, 미등록어 추정 시스템, 중국어 병음 생성 시스템을 이용하였다.



<그림 2> 운율 추정의 전반적 구조

4.1 운율 경계 추정

Li AiJun은 중국어는 운율단어(Prosodic Word), 운율구(Prosodic Phrase), 운율군(Prosodic Group)의 세 가지 운율 단위로 된 운율적 계층구조로 이루어져 있다고 했다[9]. 이는 C-ToBI의 운율 경계 표기에서 약간 차이가 나는데 본 연구에서는 MIP(minor phrase)와 MAP(major phrase)가 듣기에 분명한 차이가 없다고 판단하고

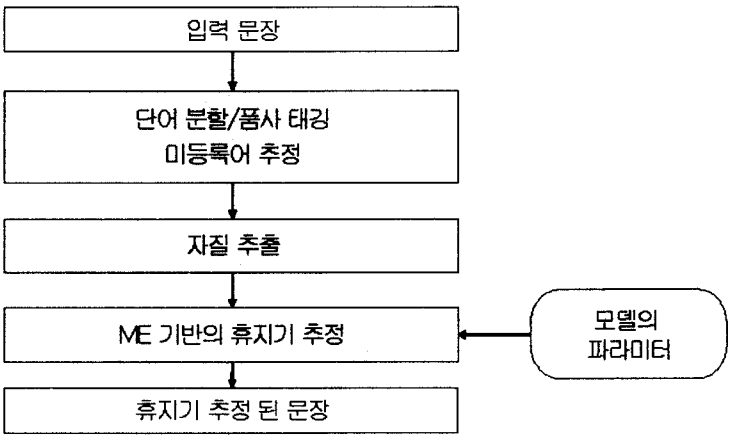
운율구로 합쳐서 네 개로 운율 경계를 표기하였다. <그림 3>은 운율 경계 표기가 된 예이다. 여기서 B_0는 단어 사이에 경계가 없을 때, B_1은 운율단어의 경계, B_2는 운율구의 경계, B_3은 운율군의 경계를 표기한다.

妹妹(n)/B_1 赶到(v)/B_0 了(u)/B_1 出事(vn)/B_0 的(u)/B_0 地点(n)/B_2

才(d)/B_1 将(p)/B_0 他(r)/B_1 送到(v)/B_0 了(u)/B_1 医院(n)/B_3 。(w)

<그림 3> 운율 경계 표기가 된 예

기존의 언어적 분석 시스템을 이용해 운율 경계 추정을 위한 두 종류의 자질들을 추출했고, <그림 4>처럼 운율 경계 추정을 위해 L-BFGS[10][11] 방법으로 예측한 조건부 ME 모델 매개변수를 사용했다.



<그림 4> 운율 경계 추정 처리

아래는 우리가 사용한 두 종류의 자질들이다.

(1) 문법적인 자질

-어휘 단어 자질(Lexical word features)

좌측(W-1), 현재(W), 우측 단어(W1)가 포함되어 있고 모두 이진 자질[2]이다.

-품사 자질(POS tag features)

운율 경계 추정에서 가장 중요한 자질이며 본 논문에서는 품사를 43개로 분류하였다. 좌측 2개(P-2, P-1), 현재(P), 우측 2개(P1, P2) 품사를 포함하며 모든 품사 자질들은 이진 자질이다.

2) 카테고리 변수와 비슷한 의미로 자질의 값이 0(안 나타남) 또는 1(나타남)만 갖는다.

(2) 수적인 자질

-길이 자질

중국어에서 단어는 기본 발음 단위이며, 구의 기본 단위이다. 그러므로 n-음절 단어는 단어의 길이가 n이고 대응되는 운율적 정보를 얻는데 사용된다. 주로 단어의 길이는 1~4로 단음절어, 2음절어, 3음절어, 4음절어이며 길이 자질은 좌측 두 개(WLen-2, WLen-1), 현재(WLen0), 우측 두 개(WLen1, WLen2)의 단어 길이를 포함한다. 길이 자질은 실수 값을 갖는 자질이며 다음과 같이 정규화를 하였다.

$$Normalized\ word\ length = \frac{Current\ word\ length}{Maximum\ word\ length} \quad (1)$$

여기서 Maximum word length는 코퍼스에 있는 단어 중 최대 길이이다.

-거리 자질

문장 시작점에서 현재 위치까지의 음절 개수인 시작 거리(dis_start)와 현재 위치에서 문장 끝까지의 음절 개수인 끝 거리(dis_end)가 포함되어 있다. 거리 자질 역시 실수 값을 갖는 자질이며 다음과 같이 정규화를 하였다.

$$Normalized\ word\ length = \frac{Distance}{Sentence\ length} \quad (2)$$

4.2 피치 악센트 추정

20세기 초, 중국의 톤과 억양 연구는 두 음성학자(Dr. Liu Fu, Dr. Chao Yuan-ren)에 의해 새로운 국면을 맞았다. Chao의 요점은 마치 큰 물결 위에 작은 물결이 탄 것처럼 문장의 억양에 의해 음절 톤의 패턴이 변할 수 있다는 것이었다. 이것은 음절 톤의 패턴과 문장의 억양 곡선과의 관계를 명확히 설명해 준다. 중국어에서 어휘 톤은 미끄러지는 피치 곡선으로 알려졌다. 고립된 곳에서 생산된 곡선들은 잘 정의되어 있고 매우 안정적이다. 반면 문맥상에서 생산된 톤의 곡선은 앞의 톤과 다음에 나오는 톤에 의존되어 많은 변화를 겪는다. 우리는 피치 악센트를 구별하기 위해 6 종류의 기호를 사용했다. <그림 5>는 피치 악센트가 표기된 예이다.

许多(m)/ B_0 电影(n)/ B_0 人(n)/ B_1 对此(d)/ B_1 也(d)/ B_1 都(d)/ B_0 有(v)/ B_0 一些(m)/ B_1 议论(v)/ B_3 .

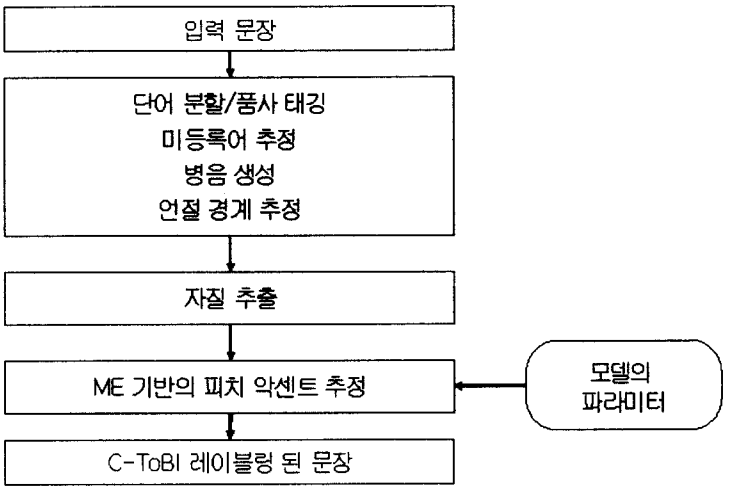
xu:3(L-H) duo1(H-H) dian4(H-H) ying3(L-L) ren2(L-L) dui4(L-L) ci3(L-L)

ye3(L-L) dou1(H-H) you3(H-L) yi4(L-) xie1(L-L) yi4(L-L) lun4(L-L) .

<그림 5> 피치 악센트가 표기된 예

우리는 중국어 병음 생성 시스템, 운율 경계 추정 시스템 등 기존의 언어분석 시스템을 이용하여 5 종류의 언어적 자질(음운, 품사, 운율 경계, 위치, 길이 자질)을 추출하였고, 조건부 ME 모델 매개변수[12]는 피치 악센트 추정을 위한 L-BFGS 방법[11]으로 예측했다(그림 6).

ME는 지수 모델들을 위한 Maximum Likelihood(ML) 훈련처럼 보여 지는데, 다른 ML 방법들이 그렇듯 훈련데이터에만 너무 맞춰지는 경향이 있다. 따라서 우리는 ME 모델들을 위해 제안된 몇몇의 방법들 중에서 Gaussian prior smoothing[13]을 채택했다. Gaussian Prior는 일반적인 ME 모델들의 smoothing을 위한 강력한 툴이고, 언어 모델에도 잘 적용할 수 있다.



<그림 6> 피치 악센트 추정시스템

우리 모델에서 사용된 자질들은 두 개의 그룹(고립된 자질 군, 복합 자질 군)으로 나눌 수 있다.

고립된(Isolated) 자질 군은 아래와 같은 uni-gram 자질들이 사용된다.

(1)음운 자질

비록 음운 곡선은 앞뒤의 톤에 의존되어 변화를 겪지만 여전히 중요한 운율 정보이고, 피치 악센트를 추정하는데 가장 널리 사용되는 것 중에 하나이다. 음운 자질은 현재 음절과 좌우측 음절의 병음, 현재 음절의 자음과 모음, 현재 음절의 톤을 갖는다. 우리가 사용한 통합 DB에는 자음 38가지, 모음 21가지, 톤 5가지, 병음 1231가지[3]가 있다.

3) 중국어 발음 체계는 1,600여개의 병음(성조(tone) 포함)으로 표기 가능한데 그 중 1000여개의 출현 빈도가 상당히 높다.

(2)문법 자질

-품사 자질

좌측품사(P-1), 현재 품사(P0), 우측 품사(P1)를 갖는다.

(3)운율 경계 자질

경계 없음(non-break), 운율단어, 운율구, 운율군을 갖는다.

(4)수적인 자질

-위치 자질

한 문장, 한 구, 각각의 단어 안에서의 음절의 위치를 갖는다. 일반적으로 문장, 구, 단어에서의 피치 곡선은 억양 패턴을 따를 것이다. 예를 들면 F0 곡선은 평서문에서 내려간다. 이는 문장, 구, 단어 안에서 음절의 위치가 운율 정보에 영향을 준다는 것을 암시한다.

-길이 자질

현재 단어 길어, 다음 단어 길이, 문장의 길이를 갖는다.

복합(Co-occurrence) 자질 군은 아래와 같이 쌍으로 사용된다.

(1)현재 품사 - 한 단어 내에서의 음절 위치.

(2)현재 운율 경계 - 한 단어 내에서의 음절 위치.

(3)현재 음절 병음 - 한 단어 내에서의 음절 위치.

(4)현재 품사 - 우측 품사.

(5)좌측 음절 병음 - 현재 음절 병음.

(6)현재 음절 병음 - 우측 음절 병음.

(7)좌측 음절 병음 - 현재 음절 병음 - 우측 음절 병음(triple feature)

4.3 피치 곡선 생성

합성모듈에서 ToBI 레이블 시스템을 이용하여 운율 톤을 생성하는 일은 입력 문장으로부터 억양 레이블을 예측하고 예측된 레이블과 다른 정보들로부터 피치 곡선을 생성하는 두 개의 부분 작업으로 구성된다. 우리는 C-ToBI 레이블들로부터 피치를 생성하기 위해서는 널리 쓰이는 선형 회귀(linear regression) 방법을 사용했다[14]. 이 방법은 레이블 타입에 대한 다른 규칙들을 필요로 하지 않으며 많은 다른 언어에도 충분히 적용 가능하다. 우리의 피치 추정 공식은 다음과 같다.

$$\tilde{P} = I + w_1 f_1 + w_2 f_2 + w_3 f_3 + \ldots\ldots + w_n f_n \qquad (3)$$

여기서 f_i는 피치에 영향을 주는 자질들이고 자질들의 가중치 w_1~w_n과 초기값 I는 선형 회귀를 통해 결정할 수 있다. 우리는 위의 공식을 모든 음절에 적용해서 피치의 추정 값을 얻었는데 이는 음성 파일로부터 추출된 피치 값들을 가지고 한

음절을 다섯 부분으로 나눈 다음 각 지점에서의 피치를 예측했다.

5. 실험 결과

5.1. 코퍼스 분석

우리의 실험은 보이스웨어에서 제공된 상용 중국어 데이터베이스를 이용하여 수행되었다. 데이터베이스는 2197개의 문장과 52,546개의 중국 한자, 25,974개의 어휘로 구성되어 있다. 데이터베이스는 품사와 병음이 표기되어 있으며, 4 종류의 운율 구조로 운율 경계가 레이블링(break-labeled)되어 있고 피치 악센트는 6 종류로 레이블링되었다. 코퍼스에서 각각의 경계 색인(break indices)의 발생 확률은 <표 2>와 같다.

<표 2> 코퍼스에서 경계 색인들의 발생 확률

B_0	B_1	B_2	B_3
36.81%	45.78%	9.58%	7.83%

코퍼스에서 피치 악센트의 발생 확률은 <표 3>과 같다. 본 실험에서는 10-fold cross validation을 했다.

<표 3> 코퍼스에서 톤들의 발생확률

H-H	H-L	L-L	L-H	H	L
16.0%	11.7%	43.4%	22.8%	2.2%	3.9%

5.2. 성능 측정법

운율 경계 추정, 피치 악센트 추정, 피치 곡선 생성에 대해 성능을 다음과 같은 방법으로 측정하였다.

먼저 **운율 경계 추정**에서는 break-correct, juncture-correct, adjusted score 세가지 측정법을 사용하였다[15][16]. 여기서 N은 모든 juncture 즉, 모든 색인의 개수고, B는 운율 경계 개수 즉, B_0를 제외한 경계 색인 개수이고, D는 삭제 에러 개수, S는 치환 에러 개수, I는 삽입 에러 개수이다. Juncture-correct식은 accuracy에 해당이 된다.

$$Break\_Correct(B\_C) = \frac{B - D - S}{B} \times 100\% \quad (4)$$

$$Juncture\_Correct(J\_C)(Accuracy) = \frac{N - D - S - I}{N} \times 100\% \quad (5)$$

$$JC = \frac{Juncture\_Correct}{100} \quad (6)$$

$$NB = \frac{N - B}{N} \quad (7)$$

$$Adjusted\_Score(A\_S) = \frac{JC - NB}{1 - NB} \quad (8)$$

피치 악센트 추정에서 사용한 성능 측정법은 Accuracy로 다음과 같이 간단히 정의된다.

$$ACC = \frac{c}{N} \times 100\% \quad (9)$$

여기서 c는 맞게 추정된 샘플의 개수고, N은 총 샘플 개수다.

피치 곡선 생성에서는 root mean squared error(RMSE) 와 상관계수(correlation coefficient) 두 가지 measurement를 사용하여 평가하였다[7].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widetilde{P}_i - P_i)^2} \quad (10)$$

$$Corr = \sum_{i=1}^{n} \frac{\frac{(\widetilde{P}_i - \widetilde{P}_{mean})(P_i - P_{mean})}{n-1}}{\frac{(\widetilde{P}_i - \widetilde{P}_{mean})^2}{n-1} \times \frac{(P_i - P_{mean})^2}{n-1}} \quad (11)$$

5.3. 운율 경계 추정 결과

우리의 ME기반 방법으로 추정된 운율 경계의 성능을 보이기 위해서 두 가지 실험을 했다. 첫 번째 실험에서 자질들의 조합으로 최상의 자질을 선택 하였고, 두 번째 실험에서는 같은 코퍼스로 품사 bi-gram 통계 모델, HMM기반 모델, 휴리스틱 규칙 기반의 에러 정정을 사용해 운율 경계를 추정한 것들과 비교하였다. 우리는 두 실험에서 4개의 경계(B_0, B_1, B_2, B_3)나 2개의 경계(B_0, B_{123})로 평가했다.

(1) 자질 선택 결과.

<표 4>와 <표 5>는 자질들의 조합을 통해 최상의 자질 선택을 한 결과로서 다음과 같은 사실을 알 수 있다. 첫째, 품사 자질은 운율 경계 추정에 기본적인 자질로 창 사이즈가 5일 때 성능이 가장 좋다. 둘째, 어휘 단어(Lexical Word)자질을 추가하면 3% 정도의 성능이 향상된다. 셋째, 길이 자질도 유용한 자질이며 창 크기를 3으로 했을 때 보다 5로 했을 때 약간 더 좋은 성능을 나타내며 정규화를 거치면 성능이 더욱 향상된다. 마지막으로 거리 자질은 정규화를 거친 후에 추가를 해야 성능이 올라간다.

<표 4> 각 자질들의 성능

Feature	Acc4[4] %
POS(-2, -1, 0, 1, 2)	81.41
Word(-1, 0, 1)	75.95
Wlen(-1, 0, 1)	66.19
Dis_start, Dis_end	63.76

<표 5> 최상 자질 선택의 결과

자질 종류	자질	B_C %	Acc4 (J_C) %	A_S	Acc2 (J_C) %
1-1	POS(-1,0,1)	79.82	77.85	0.692	83.65
1-2	POS(-2,-1,0,1,2)	81.91	81.41	0.741	86.54
1-3	POS(-3,-2,-1,0,1,2,3)	67.97	71.23	0.531	75.84
2-1	Word(0)	82.57	82.07	0.748	86.65
2-2	Word(-1,0,1)	84.98	84.19	0.762	87.80
3-1	Wlen(-1,0,1)	85.44	84.82	0.795	90.06
3-2	Wlen(-2,-1,0,1,2)	84.98	85.55	0.800	90.18
3-3	Wlen normalization	85.64	86.39	0.809	90.40
4-1	Dis_start, Dis_end	84.91	85.52	0.800	90.13
4-2	Dis_start, Dis_end normalization	85.80	86.48	0.810	90.33

4) Acc4는 4개의 경계(B_0, B_1, B_2, B_3)를 사용한 Accuracy이고 Acc2는 2개의 경계(B_0, B_{123})를 사용한 Accuracy이다.

(2) 다른 방법들과 비교

ME 기반의 방법이 품사 bi-gram의 확률적 방법이나 HMM 기반의 방법보다 좋으며 그 방법들에 휴리스틱 규칙 기반의 에러 정정을 추가하여도 ME 기반의 방법이 더 좋다는 것이 <표 6>에 보인다. 그리고 ME 기반의 방법에 휴리스틱 규칙 기반의 에러 정정을 추가하여도 성능이 거의 올라가지 않는 것을 알 수 있는데 이는 ME 기반 방법은 휴리스틱 규칙 기반에서 정한 에러가 거의 발생하지 않는다는 것을 의미한다.

<표 6> 다른 방법들과 비교

	B_C %	Acc4 (J_C)%	A_S	Acc2 (J_C)%
bi-gram	81.82	78.79	0.703	81.82
bi-gram + rules	85.80	81.02	0.732	85.86
HMM	75.62	75.05	0.623	75.62
HMM + rules	79.80	79.06	0.676	79.81
ME	85.80	86.48	0.810	90.33
ME + rules	85.92	86.55	0.812	90.54

본 실험에서 사용한 휴리스틱 규칙 기반의 에러 정정은 85개[17]로서 <그림 7>은 그 중 4개를 예제로 보여준다.

1. **if** *POS(0)=noun* **and** *WLen(0)=1* **and** *Word(1)=*标示 **then** *boundary-type*= B_0
2. **if** *POS(0)=noun* **and** *WLen(0)=1* **and** *POS(1)=noun* **and** *WLen(1)=1* **then** *boundary-type*= B_0
3. **if** *POS(0)=pronoun* **and** *POS(1)=verb* **then** *boundary-type*= B_1
4. **if** *POS(0)= auxiliary* **then** *boundary-type*= B_1

<그림 7> 운율 분절 규칙의 예

5.4. 피치 악센트 추정 결과

우리는 3개의 실험을 통해 ME기반의 피치 악센트 추정을 실험했다. 첫 번째 실험에서는 고립된 자질들의 조합을 통해 최상의 자질 선택을 보여주며, 두 번째 실험에서는 복합 자질들의 조합을 통한 최상의 자질 선택을 보여준다. 세 번째 실

험에서는 Gaussian smoothing을 거친 결과를 보여준다.

(1) 고립된 자질 선택의 결과

<표 7>은 각각의 고립된 자질의 종류에 따른 성능을 보여준다.

<표 7> 각 자질의 성능

자질	ACC %
음운 자질(-1, 0, 1)	64.67
품사 자질(-1, 0, 1)	45.63
운율 경계 자질	43.37
수적인 자질	44.15

<표 8>은 고립된 각 자질 클래스에서 최상의 자질 선택의 성능을 보여준다.

<표 8> 독립된 자질 선택의 결과

자질		Acc %
base line	톤을 기반	35.51
	각 클래스의 빈도를 기반	43.40
음운	현재의 병음	60.36
	좌, 우측의 병음	64.67
	모음, 자음, 톤	65.43
품사	현재의 품사	65.75
	좌, 우측의 품사	66.96
운율 경계		67.42
위치(음절이 문장에서의 위치)		67.68
길이	현재와 우측 단어의 길이	67.70
	문장 길이	67.73

위 실험은 피치 악센트 추정에서 고립된 자질 선택의 효과를 보여 주며 다음과 같은 결론을 유추할 수 있다.

-음운 자질은 현재 병음, 이전 병음, 다음 병음, 자음, 모음, 톤을 포함하는 가장 중요한 자질이다.

-품사 자질도 유용한 자질이며, 위 실험에서는 창 크기가 3일 때 가장 좋은 결과를 보였다.

-운율 경계 자질도 성능 향상에 도움이 된다.

-위치 자질과 길이 자질은 피치 악센트 추정에 근소한 영향만 준다.

(2) 복합 자질 선택의 결과

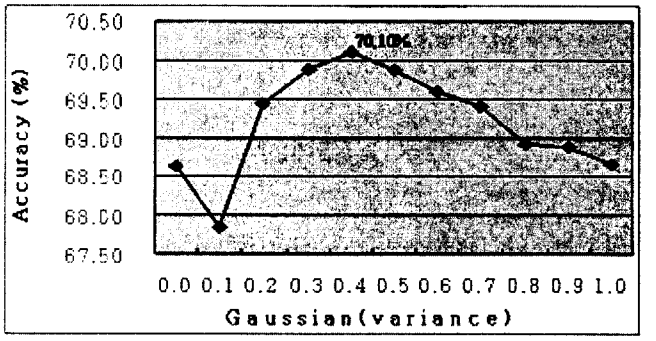
<표 9> 복합 자질 선택의 결과

자질	Acc %
현재 품사-음절의 단어 내 위치	68.60
현재 운율 경계-음절의 단어 내 위치	68.68
현재 음절의 병음-음절의 단어 내 위치	68.78
현재 품사-우측 품사	68.80
좌측 병음-현재 병음	68.88
현재 병음-우측 병음	68.81
좌측 병음-현재 병음-우측 병음	68.64

<표 9>는 복합 자질 선택의 결과이며 이 실험에서 복합 자질 또한 피치 악센트 추정에 유용하다는 걸 보여준다. 끝에 두 가지 복합 자질은 추가한 이후에 자료의 희귀성 때문에 성능이 떨어짐을 보이지만 Gaussian smoothing을 하면 [좌측 병음, 현재 병음, 우측 병음] 자질까지 추가한 것에서 최고 성능이 나온다.

(3) smoothing의 결과

<그림 8>은 [좌측 병음, 현재 병음, 우측 병음] 자질에 대한 Gaussian smoothing 결과이며 Gaussian 분산이 0.4일 때 최고 성능(70.10%)이 나타남을 보여준다. <표 10>에서는 기존의 시스템보다 많은 성능 향상이 있음이 보인다.



<그림 8> Gaussian smoothing

<표 10> 성능 향상

분류 수	Baseline %	Acc %	향상 %
6	43.40	70.10	26.70

5.5. 피치 곡선 생성 결과

우리는 각 음절 당 5개의 피치 값을 추정하고 추정된 값을 기반으로 피치를 생성했다. ESPS/Xwaves의 음성으로부터 16ms 마다 피치 값을 얻었고, 피치 열과 음(phone)열로 정렬되어 자동 생성된 파일로부터 자질 집합을 추출했다. 모델 구축을 위해 노이즈처럼 피치 값이 0인 아이템은 제거했으며, 피치 생성을 위한 선형 회귀 모델 구축에 쓰이는 자질들은 다음과 같이 7 종류를 사용하였다.

-현재 음절의 자음, 현재 음절의 모음, 현재 음절의 톤, 현재 음절의 품사,
현재 음절의 운율 경계 색인, 현재 문장에서의 음절 위치, 피치 악센트

피치 곡선 추정의 방법은 각 음절로부터 5개의 피치를 추출한 뒤, 각 피치를 위한 선형 회귀 모델을 구축해서 5개의 모델을 얻었다. 이 5개의 모델에서 추정된 응답을 벡터 형태로 얻어내고, 그것들의 평균과 각 요소로 RMSE와 상관계수를 계산했다. <표 11>에서는 C-ToBI를 사용 했을 때 와 C-ToBI를 사용하지 안 했을 때의 결과를 비교하여 C-ToBI 기반 추정 모델을 사용하면 피치 생성 성능이 개선됨을 보인다. 여기서 C-ToBI를 사용 안 했을 때는 피치 곡선 추정 공식 (3)에서 C-ToBI 자질을 빼고 추정한 것을 의미한다.

<표 11> 피치 생성의 결과

	C-ToBI 자질을 사용 안 했을 때	C-ToBI 자질을 사용 했을 때	성능 향상
RMSE	47.920	40.552	-7.368
Correlation coefficient	0.531	0.621	+0.090

6. 결론

본 논문은 중국어에서 C-ToBI 시스템기반으로 운율 요소(운율 경계와 피치 곡선)를 생성하는 방법을 조건부 최대 엔트로피(ME)모델을 사용하여 제안했다. ME 기반으로 운율 경계와 피치 곡선을 추정하기 위해 언어적, 음성적 자질들을 분석하여 최상의 후보들을 찾았으며, 완전 자동 톤 레이블링 시스템을 이용해 피치 악센트가 추정된 코퍼라를 생성했다. 그 결과 우리가 제안한 모델로부터 좋은 성능을 보장하는 최상의 자질 집합을 얻었다. 게다가 보통 중국어 운율 경계와 피치 곡선에 관련된 자질들은 서로 의존적인데 우리가 제안한 ME모델의 자질 선택은 다양한 의존성으로부터 자질들을 독립되게 만들기 때문에 다른 기계 학습 모델보다 호환성이 높다. 우리의 결과는 C-ToBI를 이용한 운율 생성 성능이 C-ToBI를 사용하지 않았을 때보다 좋음을 보여준다. 앞으로 성능을 더 개선하기 위해서는 보

다 다양한 자질들과 보다 좋은 smoothing 기법을 연구해야 하겠다.

감사의 글

본 연구는 과학재단 특정기초(R01-2003-000-10181-0) 지원을 받아 수행되었음. 우리는 중국어 음성합성 데이터베이스를 제공해준 (주)보이스웨어에 깊은 감사를 드립니다.

참고 문헌

[1] A. Steven, "Chunks and dependencies: bringing processing evidence to bear on syntax," *Computational Linguistics and Foundations of Linguistic Theory,* CSLI, pp.145-164, 1995.

[2] Z. Ying, X. Shi, "An RNN-based algorithm to detect prosodic phrase for Chinese TTS," *in Proc. of ICASSP*, pp.809-812, 2001.

[3] Q. Shi, X. Ma, et al., "Statistic prosody structure prediction," *in Proc. of IEEE workshop on speech synthesis,* pp.155-158, 2002.

[4] Z. Sheng, T. J. Hua, C. L. Hong, "Learning rules for Chinese prosodic phrase prediction," *COING*, pp.79-88, 2002.

[5] X. Sun, "Pitch accent prediction using ensemble machine learning," *in Proc. of ICSLP,* pp.953-956, 2002.

[6] M. L. Gregory, Y. Altun, "Using conditional random fields to predict pitch accents in conversational speech," *ACL,* pp.677-683, 2004.

[7] K. Dusterhoff, A. W. Black, "Generating F0 contours for speech synthesis using the tilt intonation theory," *The ESCA workshop on intonation: Theory, Models and Application,* pp.107-110, 1997.

[8] J. Lee, B. Kim, G. G. Lee, "Automatic corpus-based tone and break-index prediction using K-TOBI representation," *ACM transactions on asian language information processing (TALIP),* Vol. 1, No. 3, pp.207-224, 2002.

[9] L. AiJun "Chinese prosody and prosodic labeling of spontaneous speech," *Speech Prosody,* pp.39-46, 2002.

[10] Z. Sheng, T. J. hua, J. D. Ling "Chinese prosodic phrasing with extended features," *in Proc. of ICASSP,* Vol 1, pp.492-495, 2003.

[11] R. Malouf, "A comparison of algorithms for maximum entropy parameter estimation," *In proceeding of CoNLL-2002,* pp.49-55, 2002.

[12] A. L. Berger, S. A. D. Pietra, V. J. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics,* Vol.22, No.1, pp.39-72, 1996.

[13] S. F. Chen, R. Rosenfeld, "A Gaussian prior for smoothing maximum entropy models,"

Technical Report CMU-CS-99-108, 1999.

[14] A. W. Black, A. J. Hunt, "Generating F0 contours from ToBI lables using linear regression," *In Proc. of ICSLP,* pp.49-55, 1996.

[15] P. Taylor, A. W. Black, "Assigning phrase breaks form part-of-speech sequence," *Computer Speech and Language,* Vol. 12, No. 2, pp.99-117, 1998.

[16] E. Sanders, "Using probabilistic methods to predict phrase boundaries for a text-to-speech system," Master's thesis, University of Nijmegen, 1995.

[17] J. Cao, "Syntactic and lexical constraint in prosodic segmentation and grouping," *Speech Prosody,* pp.203-206, 2002.

접수일자: 2005년 2월 10일
게재결정: 2005년 3월 15일

▶김승원(Seungwon Kim)
주소: 790-784 경상북도 포항시 남구 효자동 산 31번지 포항공과대학교
소속: 포항공과대학교 컴퓨터공학과
전화: 054) 279-5581
E-mail: rockzja@postech.ac.kr

▶정옥(Yu Zheng)
주소: 790-784 경상북도 포항시 남구 효자동 산 31번지 포항공과대학교
소속: 포항공과대학교 컴퓨터공학과
전화: 054) 279-5581
E-mail: zhengyu@postech.ac.kr

▶이근배(Gary Geunbae Lee)
주소: 790-784 경상북도 포항시 남구 효자동 산 31번지 포항공과대학교
소속: 포항공과대학교 컴퓨터공학과
전화: 054) 279-5581
E-mail: gblee@postech.ac.kr

▶ 김병창(Byeongchang Kim)
주소: 717-702 경상북도 경산시 하양읍 금락1리 330번지 대구가톨릭대학교
소속: 대구가톨릭대학교 컴퓨터정보통신공학부
전화: 053) 850-2718
E-mail: bckim@cu.ac.kr