벡터 양자화를 이용한 한국어 억양 곡선 생성

안혜선° 김형순
부산대학교 인지과학 협동과정, 부산대학교 전자공학과
{dodobird, kimhs}@hyowon.pusan.ac.kr

Generation of Korean Intonation using Vector Quantization

Hye-Sun An° Hyung-Soon Kim
Dept. of Interdisciplinary Research Program of Cognitive Science,
Dept. of Electronics Engineering, Pusan National University

요 약

본 논문에서는 text-to-speech 시스템에서 사용할 억양 모델을 위해 벡터 양자화(vector quantization) 방식을 이용한다. 어절 경계강도(break index)는 세단계로 분류하였고, CART(Classification And Regression Tree)를 사용하여 어절 경계강도의 예측 규칙을 생성하였다. 예측된 어절 경계강도를 바탕으로 운율구를 예측하였으며 운율구는 다섯 개의 억양 패턴으로 분류하였다. 하나의 운율구는 정점(peak)의 시간축, 주파수축 값과 이를 기준으로 한 앞, 뒤 기울기를 추출하여 네 개의 파라미터로 단순화하였다. 운율구에 대해서 먼저 운율구가 문장의 끝일 경우와 아닐 경우로 분류하고, 억양 패턴 다섯 개로 분류하여, 모두 10개의 운율구 set으로 나누었다. 그리고 네 개의 파라미터를 가지고 있는 운율구의 억양 패턴을 벡터 양자화 방식을 이용하여 분류(clusteing)하였다. 운율의 변화가 두드러지는 조사와 어미는 12 point의 기본주파수 값을 추출하고 벡터 양자화하였다. 운율구와 조사, 어미의 codebook index는 문장에 대한 특징 변수 값을 추출하고 CART를 사용하여 예측하였다. 합성할 때에는 입력 text에 대해서 운율구의 억양 파라미터를 추정한 다음, 조사와 어미의 12 point 기본주파수 값을 추정하여 전체 억양 곡선을 생성하였고 본 연구실에서 제작한 음성합성기를 통해 합성하였다.

1. 서 론

Man-machine interface의 핵심 기술 중의 하나인 문장-음성 변환(text-to-speech(TTS) conversion) 기술은 사용자가 입력한 문장을 기계가 분석한 후 미리 저장된 음성의 기본 단위들로부터 음성신호를 합성하는 것이다. 본 논문에서는 문장-음성 변환 기술의 자연성 향상을 위해서 벡터 양자화 방식을 이용한 억양(intonation) 모델을 제안한다.

억양은 운율(prosody)의 한 요소로서, 사람이 발성 때 일어나는 음의 높이(pitch) 변화, 즉 성대의 진동에 의해서 생기는 음조의 높이의 변화를 말한다. 억양의 물리적 의미는 기본주파수(Fundamental frequency, F0)의 변화에 해당한다. 억양의 특성을 기술하기 위한 연구에는 ToBI[1], RFC model[2], tilt model[3], maximum-based model[4], parametric intonation event (PaIntE) parameter[5] 등이 있으며, 억양의 생성 방법론에는 CART[6], linear regression[7], sums-of-product model[8], rule-based approach[9], neural networks[10] 등이 있다.

본 논문에서는 억양 곡선을 몇 가지의 파라미터로 단순하게 나타낼 수 있다는 점에 착안하여 억양 곡선을 maximum-based model 파라미터[4]로 단순화하였다. 그리고 입력 문장의 어절 경계강도 예측, 운율구의 억양 곡선 예측과 조사, 어미의 억양 곡선 예측을 위하여 CART 방식을 사용하였다. 운율구의 억양 곡선 패턴을 분류하기 위하여 벡터 양자화 방식을 사용하였고, 운율의 변화가 두드러지는 조사, 어미의 기본주파수 분류에 대해서도 벡터 양자화 방식을 사용하였다.

2. 음성 데이터베이스 및 운율 레이블링

2.1. 음성 데이터베이스

운율 생성을 위한 데이터베이스는 '청산유수' [11] 음성 데이터베이스 중에서 남성 화자의 530개 문장을 선정하였다. '청산유수' 데이터베이스는 현직 아나운서를 통해 방음 부스에서 녹음하였으며, 발성 목록은 뉴스, 일기예보, 교통안내, 신문사설, 소설, 교과서 등에서 골고루 추출한 것이다.

530개 문장 중에서 360개의 문장은 훈련용 문장으로, 170개의 문장은 검증용 문장으로 사용하였다. 훈련 데이터의 운율구의 개수는 1079개, 어절 개수는 3048개로 구성되어 있다. 검증 문장의 운율구 개수는 518, 어절 개수는 1395개로 구성되어 있다.

표 1. 음성 DB의 구성

	훈련용 데이터	검증용 데이터
문장 개수	360	170
운율구 개수	1079	518
어절 개수	3048	1395

2.2. 운율 레이블링

운율 생성을 위한 운율 레이블링은 Entropic사의 ESPS/Xwaves+ 프로그램을 사용하였다.

어절 경계강도는 '끊어읽기 색인', '운율 경계강도' 라고도 하며, 발화된 음성을 청취할 때 느끼는 어절 간의 운율적 이질감으로서 객관적이고 물리적인 현상이 아니라 지각적이고 심리 음향적인 현상이라고 할 수 있다. 이 경계강도에 따라서 휴지기의 유무와 길이, 경계 음절의 길이 및 억양 변화가 특징적으로 나타나기 때문에 어절 경계강도를 설정하고 예측하는 것은 합성음의 자연스러움에 매우 큰 영향을 미친다.

본 논문에서 어절 경계강도는 쉼이 없는 어절 경계(BI-0), 쉼 이 있고 구의 마지막 음절이 길어지거나 억양 변화가 경계로 지각되는 어절의 경계(BI-1), 문장의 마지막 어절(BI-2), 세단계로 분류하였다. 표 2 는 어절 경계강도의 정의와 개수이다.

표 2. 어절 경계강도의 정의와 개수

단계	정의	음성 DB에서의 개수	
		훈련 데이터	검증 데이터
BI-0	쉼이 없는 어절 경계	1969	877
BI-1	운율구 경계	719	348
BI-2	문장 마지막 어절	360	170

운율구는 표 3 과 같이 다섯 단계로 나누었다. 운율구는 톤 패턴(tonal pattern)에 따라 'LHL' , 'HL' , 'LH' , 세 범주로 나누었다. 그리고 'LHL' 은 'L' 톤과 'H' 톤의 높이 차이가 클 경우(LHL1)와 높이 차이가 작을 경우(LHL2)로 나누었다. 마찬가지로, 'HL' 톤도 'H' 톤과 'L' 톤의 높이 차이가 클 경우(HL1)와 작을 경우(HL2)로 나누었다.

표 3. 운율구의 정의와 개수

단계	정의	음성 DB에서의 개수	
		훈련 데이터	검증 데이터
P-1	LHL1	368	169
P-2	LHL2	487	244
P-3	HL1	84	49
P-4	HL2	119	48
P-5	LH	21	8

3. 어절 경계강도 예측

어절 경계강도는 표 2와 같이 'BI-0' , 'BI-1' , 'BI-2' 으로 나누어 예측하였다. 예측 방법은 Tree 기반 모델링 기법 중 하나인 CART를 이용하였다. CART는 패턴분석이나 회귀 분석을 통해 운율구의 설정, 휴지 길이 추정 등의 운율 구조 생성에 매우 유용하게 사용되고 있는 통계적 방법이다.

예측 결과는 표 4 와 5 에서 예측된 class의 개수와 그것을 백분율로 환산한 값을 괄호 안에 나타내었다. 문장의 마지막 어절인 'BI-2' 을 제외하고, 훈련 데이터에서는 16.51%의 예측 오류율을 나타내었으며, 검증 데이터에서는 16.91%의 예측 오류율을 나타내었다. 결정 트리에서 종결 노드의 개수는 18개였다.

표 4. 훈련 데이터의 어절 경계강도 예측 결과

훈련 데이터	Predicted Class		
Actual Class	BI-0	BI-1	BI-2
BI-0	1601 (81.31%)	368 (21.69%)	0 (0%)
BI-1	103 (17.39%)	616 (85.67%)	0 (0%)
BI-2	0 (0%)	0 (0%)	360 (100%)

표 5. 검증 데이터의 어절 경계강도 예측 결과

검증 데이터	Predicted Class		
Actual Class	BI-0	BI-1	BI-2
BI-0	729 (83.12%)	148 (16.88%)	0 (0%)
BI-1	59 (16.95%)	289 (83.05%)	0 (0%)
BI-2	0 (0%)	0 (0%)	170 (100%)

4. 운율구 예측

운율구는 어절 경계강도를 기준으로 설정하였다. 마지막 어절의 어절 경계강도가 '1' 인 어절과 '2' 인 어절에 운율구를 할당하였다.

4.1. 운율구 기본주파수 값 추출

기본주파수 값은 Entropic사의 ESPS/Xwaves+를 사용하여 추출하였다. 먼저 5ms 간격으로 기본주파수를 추출하여 5 point median filtering을 거친 후 다시 7 point median filtering을 통해 피치 값을 smoothing하였다. 그리고 무성음 구간은 선형 보간법을 통하여 연결하였다.

4.2. 운율구 파라미터 추출

훈련 데이터의 모든 운율구의 기본주파수 값을 추출한 후 기본주파수 값을 이용하여 운율구 단위로 maximum-based model 파라미터를 추출하였다.

이 모델은 억양 곡선에서 인지적, 음향적으로 가장 큰 영향을 미치는 것은 돋들림(prominence)이며, 억양 곡선에서 돋들림을 나타내는 정점(peak)을 정의할 수 있는 파라미터들을 정한다. 따라서 본 논문은 하나의 운율구의 억양 곡선을 maximum-based model에 따라 다음과 같은 네 가지 파라미터로 단순화하였다.

그림 1와 같이, 첫번째 파라미터는 peak delay로서 액센트 음절 모음 시작에서 정점(peak)까지의 상대적 시간으로

정의한다. 즉, 정점이 모음의 시작 이후에 오면 양(+)의 값 값을 가지며, 모음 시작 이전에 오면 음(-)의 값을 가진다.

두번째 파라미터는 정점의 크기(amplitude)로서 상위값과 기저값 사이의 비율(%)값으로 나타낸다. 상위값과 기저값은 화자의 피치영역을 나타내는 것으로써 억양곡선의 임계값 역할을 한다. 상위값은 158 Hz이며, 기저값은 77 Hz로 정하였으며, 이 값들은 남성 화자의 평균 피치 값을 계산하여 구하였다.

나머지 두 파라미터는 앞 기울기(left slope)와 뒤 기울기(right slope)로서 정점에서의 억양구의 앞,뒤 음절들에 대한 기울기 값이다.

음성 DB에서 자동적으로 네 개의 파라미터를 추출한 후, 표 6 과 같이 어절 경계강도와 운율구 범주에 따라 10개의 set으로 운율구를 나누었다.

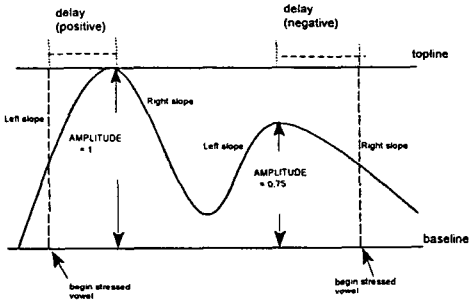

그림 1. Maximum-based model에서의 peak 파라미터

표 6. 운율구 set

이름	설명		개수
	구의 마지막 어절 경계강도	운율구 범주	
B1P1	B1-1	P-1	311
B1P2	B1-1	P-2	284
B1P3	B1-1	P-3	50
B1P4	B1-1	P-4	54
B1P5	B1-1	P-5	20
B2P1	B1-2	P-1	57
B2P2	B1-2	P-2	203
B2P3	B1-2	P-3	34
B2P4	B1-2	P-4	65
B2P5	B1-2	P-5	1

4.3. 운율구의 벡터 양자화(Vector Quantization)

벡터 양자화는 음성 인식, 음성 압축, 화상 처리 등에서 널리 사용되며, 입력 벡터를 대표 패턴이 저장된 코드북으로부터 이에 대응되는 인덱스로 표현하는 방법이다[12][13].

음성 합성 분야에서도 억양 곡선을 생성하기 위해서 벡터 양자화 방식을 사용한 사례가 있다[5][14]. 김성환과 김진영은 음절 단위로 기본주파수 개수를 벡터로 사용하여 벡터 양자화를 시행하였고[14], Mohler와 Cokie는 음절 단위로 Parametric intonation event(PaIntE) parameter를 추출하여 벡터 양자화하였다[5].

본 논문에서는 네 개의 파라미터로 구성된 10개의 운율구 set에 대해서 벡터 양자화를 하였다. 분류 방법은 K-means 알고리즘을 적용하였고 거리 척도로는 정규화된 유클리드 거리 방법을 사용하였다. 'B2P5' 운율구는 개수가 1개이기 때문에 벡터 양자화를 하지 않았다.

각 운율구 set 별로 벡터 양자화를 실시한 후, CART 방법을 사용하여 대표 패턴을 예측하였다. CART 훈련을 한 결과, 10개의 운율구 범주의 codebook 크기는 다음과 같다.

표 9. 10개 운율구 범주의 Codebook 크기

운율구 범주	Codebook 크기
B1P1	4
B1P2	4
B1P3	2
B1P4	4
B1P5	2
B2P1	4
B2P2	8
B2P3	2
B2P4	4
B2P5	1
합	35

4.4. 조사, 어미의 기본주파수 예측

한국어는 조사, 어미가 발달한 굴절어이다. 조사, 어미는 형태론적으로 많은 정보를 담고 있으며 의미론, 화용론적으로도 많은 정보를 전달하기 위해서 운율의 변화가 두드러진다. 따라서, 조사, 어미에서 피치 변화, 휴지 길이의 변화, 마지막 음절의 장음화 등의 운율 변화가 많이 일어난다. 그래서 본 논문에서는 각 조사, 어미 별로 벡터 양자화를 하여 같은 조사, 어미라도 입력 텍스트의 특징에 따라 다른 값을 예측할 수 있도록 하였다.

조사, 어미 별로 12 point 기본주파수를 추출한 후, 남성 화자의 상위값과 기저값 사이의 비율값으로 나타내었다. 그리고 각 조사, 어미 별로 12차 벡터 양자화를 하였다. 분류 알고리즘은 운율구에 대한 벡터 양자화할 때와 마찬가지로 K-means 알고리즘을 사용하였고 거리 척도는 유클리드 거리 방법을 사용하였다. 그리고 CART를 사용하여 벡터 양자화의 결과인 codebook index를 예측하였다.

4.5. 최종 억양 곡선 생성

최종 억양 곡선 생성 과정은 두 단계로 나누어진다. 첫번째 단계에서 운율구의 억양 모양(shape)을 만들어준 후, 두번째 단계에서 조사, 어미의 기본주파수 값을 더해준다. 즉, 35개의 운율구 대표 패턴 중에서 입력 텍스트의 특징 변수에 따라 하나를 선택한다. 대표 패턴의 네 가지 파라미터 값으로 운율구의 억양 모양을 모델링한다. 그리고 예측한 조사, 어미의 대표 패턴을 더한다.

이렇게 만들어진 억양 모델을 본 연구실에서 개발한 '청산유수' [11] 시스템에 결합하였다. 합성음의 자연성을 평가하기 위해 Peak 파라미터와 피치 검색 테이블을 이용한 억양 모델[15]로 구현된 '청산유수' 와 본 논문에서 제안하는 억양 모델로 구현된 '청산유수' 의 합성음을 비공식적인 청취평가로 비교한 결과, 본 논문에서 제안하는 방식

이 더 자연스럽다는 결론을 얻을 수 있었다.

5. 결 론

본 논문에서는 text-to-speech 시스템에서 사용할 억양 모델을 위하여 벡터 양자화 방식을 이용하였다. 운율구에 대해서는 maximum-based model 파라미터를 추출하여 벡터 양자화를 하였고, 조사, 어미에 대해서는 12 point의 기본 주파수 값을 추출하여 벡터 양자화하였다. 벡터 양자화의 결과인 대표 패턴은 음성 DB의 특징 변수 값을 자동으로 구한 다음 CART 방법으로 예측하였다.

운율구의 억양 패턴을 분류할 때 정규화된 유클리드 거리 척도를 사용하였는데, 향후 보다 정교하게 억양 패턴을 분류할 수 있는 유클리드 거리 방식을 적용해서 운율구의 벡터 양자화를 할 계획이다.

6. 참고문헌

[1] S.-A. Jun, "K-ToBI (Korean ToBI) labelling conventions (version 3.0)," *Speech science*, Vol. 7, No. 1, 2000.

[2] P. Taylor, "The rise/fall/connection model of intonation," *Speech communication* Vol. 15, pp.169-186, 1994.

[3] P. Taylor, "Analysis and synthesis of intonation using the tilt model," *Draft Journal Paper on Tilt Model*, 1997.

[4] B. Heuft and T. Portele, "Synthesizing prosody : a prominence-based approach," *in Proc. of ICSLP' 96*, pp.2387-2390, 1996.

[5] G. Mohler and A. Conkie, "Parametric modeling of intonation using vector quantization," *in Proc. of 3rd ESCA Workshop on Speech Synthesis*, 1998

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

[7] A. W. Black and A. J. Hunt, "Generating F0 contours from ToBI labels using linear regression," *in Proc. of ICSLP' 96*, pp.1385-1388, 1996.

[8] J.P.H. van Santen, "Analyzing N-way tables with sums-of-products models," *J. Math. Psychology*, Vol. 37, No. 3, pp.327-371, 1993.

[9] M. Jilka, G. Mohler and G. Dogil, "Rules for the generation of ToBI-based American English intonation," *Speech communication* Vol. 28, pp.83-108, 1999.

[10] C. Traber, "F0 generation with a database of natural F0 patterns and with a neural network," *in Talking Machines: Theories, Models, Designs* (G. Bailly, C. Benoit, and T.R. Sawallis, eds.), pp.287-304, Elsevier science, 1992.

[11] 김형순 외, *PC용 TEXT-TO-SPEECH 시스템 개발에 관한 연구*, 산업자원부 최종보고서, 1999.

[12] 오영환, *음성언어정보처리*, 홍릉과학출판사, 1998.

[13] L. Rabiner and B. Juang, *Fundamentals of speech recognition*, Prentice-Hall International, Inc, 1993

[14] 김성환, 김진영, "Efficient Method of Establishing Words Tone Dictionary for Korean TTS system," *in Proc. of Eurospeech' 97*, pp.247-250, 1997.

[15] 장석복, 김형순, "Peak 파라미터와 피치 검색테이블을 이용한 억양 생성방식 연구," 제 11회 한글 및 한국어 정보처리 학술대회, pp.184-190, 1999.