코퍼스 기반 한국어 합성기의 억양 구현 방안*

A Method of Intonation Modeling for Corpus-Based Korean Speech Synthesizer

김 진 영 · 박 상 언 · 엄 기 완** · 최 승 호***
Jin-Young Kim · Sang-Eon Park · Ki-Wan Eom · Seung-Ho Choi

ABSTRACT

This paper describes a multi-step method of intonation modeling for corpus-based Korean speech synthesizer. We selected 1833 sentences considering various syntactic structures and built a corresponding speech corpus uttered by a female announcer. We detected the pitch using laryngograph signals and manually marked the prosodic boundaries on recorded speech, and carried out the tagging of part-of-speech and syntactic analysis on the text. The detected pitch was separated into 3 frequency bands of low, mid, high frequency components which correspond to the baseline, the word tone, and the syllable tone. We predicted them using the CART method and the Viterbi search algorithm with a word-tone-dictionary. In the collected spoken sentences, 1500 sentences were trained and 333 sentences were tested. In the layer of word tone modeling, we compared two methods. One is to predict the word tone corresponding to the mid-frequency components directly and the other is to predict it by multiplying the ratio of the word tone to the baseline by the baseline. The former method resulted in a mean error of 12.37 Hz and the latter in one of 12.41 Hz, similar to each other. In the layer of syllable tone modeling, it resulted in a mean error rate less than 8.3% comparing with the mean pitch, 193.56 Hz of the announcer, so its performance was relatively good.

Keywords : speech synthesis, intonation modeling

I. 서 론

최근에 이르러 음성을 사용한 인간-기계간의 대화(Man-Machine Interface, MMI)를

* 본 연구는 1998년 한국학술진흥재단의 학제간연구 지원에 의하여 이루어짐.
** 전남대학교 전자공학과
*** 동신대학교 정보통신공학과

이용한 여러 서비스가 개발되어 상용화되고 있으며, 그 핵심 기술인 음성합성과 음성인식에 대한 연구가 활발히 진행되고 있다. 특히 음성합성 분야의 경우, 상당히 이해도가 높은 합성기들이 이미 개발되어, 일부 기업에서는 web site상에서 서비스를 제공하기도 하고, LG, 삼성 등의 각 기업에서도 개발된 음성합성기를 시판하고자 하는 상황에 있다[1-5]. 그러나 아직까지 시판되어 상용화되고 있는 합성기의 수는 미미한 실정이다. 이렇듯 음성합성기에 대한 연구 및 개발이 활발하면서도 그 활용이 매우 제한되어 있는 까닭은 합성음의 부자연성으로 인한 거부감 때문이다. 세계적으로 볼 때 운율에 대한 연구는 음성합성에 대한 연구가 시작된 1970년대부터 꾸준히 연구되어 왔는데, 주로 일본어 및 영어권을 중심으로 진행되어 왔으며 상당히 괄목할 만한 성과를 거두어 왔다[6-16]. 일본의 경우, Fujisaki 모델이라는 일본어 억양발생 모델은 세계적으로 인정받은 바 있으며, ATRL (Advanced Telecommunication Research Labs)음성 합성시스템은 상당한 자연스러움을 보이고 있다고 한다. 미국의 경우, MIT의 DECtalk, Berkley Speech Tech.의 BST system, 그리고 Apple, IBM 등의 여러 연구소 및 기업에서 연구가 활발히 진행되어 고품질 음성합성기가 상용화되고 있다. 다행히 국내에서도 최근에 한국어 운율현상을 규명하고자 하는 연구노력이 활발히 수행되고 있다. 특히, ETRI, KAIST, 삼성, LG 그리고 몇몇 대학에서 연구가 진행중이다[17-21]. 또한 연구의 방향도 과거의 top-down 방식을 벗어나, 음성 코퍼스 중심의 bottom-up 방식의 연구가 진행되고 있으며, ETRI의 연구결과가 대표적이라 할 수 있다. 그러나 한국어 운율에 대한 연구는 아직 많은 연구가 집중되어야 할 것이다.

본 논문은 음성 합성기 개발에 한국어의 특징적인 운율 정보를 추가함으로써 보다 자연스러운 합성음을 얻을 것을 목표로 하였다. 합성음의 자연성 향상을 위한 운율 처리 문제인 음의 장단, 강약과 아울러 중요한 요소가 바로 자연스러운 억양의 구현이다. 이는 피치의 구현과 직접적으로 연관되는 문제인데, 이를 제대로 해결하지 못한 경우 합성음의 억양의 급격한 굴곡현상 등으로 인해 청자로 하여금 인공적이고 거북한 느낌을 갖게 한다.

본 논문에서는 억양의 발생모델을 3단계, 즉 기저선을 결정하는 단계, 어절의 평균 톤을 결정하는 단계, 음절의 평균 톤과 톤의 상승하강을 결정하는 단계로 나눈다. 억양을 이처럼 구분짓기 위해 Hanning 윈도우를 사용한 필터링을 통해 구해진 피치를 3개의 주파수 영역으로 분리하였다. 어절의 평균 톤을 구현하기 위한 방법으로 기저선과 기저선에 대한 중간 주파수 영역 피치의 비를 각각 학습시킨 후 그 곱으로 예측하는 2단계 방법과 어절의 평균 톤에 해당하는 중간 주파수 영역의 피치를 학습시킨 후 예측하는 1단계 방법을 CART를 사용하여 비교 분석하였으며, 어절의 평 음절 톤에 해당하는 고주파 영역에서는 음절수에 따른 사전을 구축하고 각 파라미터에 대해 최소거리에 있는 N개 후보 단어를 대상으로 Viterbi 탐색 알고리즘을 사용하여 피치를 예측하였다.

II. 음성 코퍼스의 구축

본 논문에서 사용된 문장은 한국어의 다양한 구문 및 음운 구조를 포함하도록 선택되었으며, 여기에는 서술문, 의문문, 구문 등 어미에 따른 문의 종류와 단문, 중문, 복문 등 여러 가지의 문법구조가 포함되었다. 본 논문에서는 이러한 목적으로 여러 가능한 텍스트(뉴스, 논설문, 소설, 수필 등)에서 1,833문장을 발췌하였다. 그리고 표준말을 사용하는 대학 방송국의 아나운서인 여성화자가 발췌된 1,833개의 문장을 보통 속도로 발성한 것을 총 16시간 동안 녹음하였다. 녹음시 Laryngograph(Lx) 신호와 음성 신호를 동시에 2채널로 녹음하는 형식으로 작업을 하였는데, 채널 하나는 음성신호를 녹음하기 위하여 사용되었으며, 다른 하나는 성대의 떨림 신호를 잡을 수 있는 Lx 신호를 녹음하였다. Lx 신호를 녹음하는 이유는 Lx 신호를 이용할 때 성문의 닫힘과 열림의 구분이 확연하여 자동으로 피치 마킹을 하는데 유용하기 때문이다. 사용된 녹음기는 SONY의 DAT(Digital Audio Tape-Recoder) TCD-D10 PRO II이며 콘덴서 마이크로폰(AKG C535EB)을 사용하였다. 이들 신호는 표본화율 8 kHz, 16 bits로 A/D 변환되었으며, 음성 데이터의 태깅과 피치 분석 등의 과정은 자체적으로 개발한 Tool을 사용하였다.

녹음된 코퍼스의 규모는 총 1,833문장(22,352어절)로, 여성 화자 1인에 의해 발음되었으며, 학습용으로 1,500문장(18,356)어절이, 테스트용으로 333문장(3,996어절)이 각각 사용되었다.

예측 파라미터는 해당 어절의 품사 및 선행/후행 어절의 품사, RD, ToRight, 운율 경계 강도, 해당 어절의 문장 전체에서의 위치 및 휴지기로부터의 위치 등을 CART를 실행하여 변수 중요도를 검증한 후 선별적으로 적용하였다.

III. 파라미터의 설정

본 논문에서 사용된 파라미터는 운율 경계 강도, 어절의 품사 정보, RD, ToRight, 문장 내 어절의 위치 등이며, 각 파라미터의 설정은 다음과 같다.

표 69. 피치 예측 파라미터의 종류

변수명	의 미
BI	운율 경계 강도
LPOS	관측어절의 좌 품사
POS	관측 어절의 품사
RPOS	관측 어절의 우 품사
RD	Right Dependency
ToRight	지배소까지의 거리
LOC	문장내에서의 상대적 위치
LOC_SP	휴지기로부터의 상대적 위치

3.1 운율 경계 강도의 결정

운율 경계 강도는 발화된 음성을 청취할 때 사람이 느끼는 어절간의 운율적 이질감으로서 객관적인 판단에 의한 값이라기보다는 심리음향학적 파라미터라고 할 수 있다. 따라서 운율 경계 강도와 운율 정보는 상호 연관성은 있으나, 반드시 일치하는 것은 아니다. 본 논문에서는 청취 실험을 통하여 운율 경계 강도를 결정하였다. 구축된 1,833문장을 들려주고 청취자들이 0~2 사이의 값 중에서 하나를 결정하도록 하였다. 청취 실험에 참가한 인원은 모두 3명이다.

청취 실험을 위해 정의한 각 운율 경계 강도의 의미는 다음과 같다[22].

표 70. 운율 경계 강도의 정의

운율경계강도	설 명
0	어절간에 경계가 전혀 느껴지지 않는다.
1	어절간에 약한 경계가 느껴진다.
2	어절간에 강한 경계가 느껴진다.

3.2 품사의 분류

본 논문에서 사용된 품사 집합은 총 18품사로서 김선미의 분류[23]를 따르며, 품사 집합은 다음과 같다.

표 71. 사용된 품사 집합의 정의

기 호	품사 구조
NV	명사류 + 부사격 조사
NO	명사류 + 목적격 조사
VT	동사류 어간 + 주, 보조적 연결 어미
VS	동사류 어간 + 종속 연결 어미
E	부사어
NS	명사류 + 주격 조사
VC	동사류 어간 + 종결 어미
VN	동사류 어간 + 명사형 어미
NG	명사류 + 속격 조사
VA	동사류 어간 + 관형사형 어미
VQ	동사류 어간 + 인용형 조사
D	관형어
B	의존 명사
VD	동사류 어간 + 종결 어미
NJ	명사류 + 연결 조사
AB	접속사
NM	명사류 + 보조사
N	명사

위의 품사 분류에 따라 각각의 어절에 품사 정보를 태깅하였으며, 관측 어절의 선행·후행 품사를 동시에 사용하였다.

3.3 구문 분석

구문 분석은 어절들간의 관계를 지배소와 의존소로 표현하는 의존 문법으로 표현한다. 의존 문법에 의해 문장 내 모든 어절들은 자신의 지배소를 하나씩 갖게 되며, 문장의 마지막 어절은 항상 자신을 지배소로 갖는다. 또한 한국어는 영어와 달리 지배소가 항상 의존소 뒤에 위치하는 것이 관측되며, 이를 지배소 후위법칙(governor post-positioning property)라고 부른다[24].

RD는 문장의 수식 관계 트리 구조의 Depth를 나타내는 파라미터로서 주어진 어절에 대해서 자기 자신의 Depth와 다음 어절의 Depth의 차이에 1을 더해준 값을 말한다. ToRight는 문장 성분 연결에서 관측어절로부터 지배소까지의 거리를 나타낸다.

다음의 그림 1과 표 4는 문장 "동물의 세계를 지배하는 유일한 힘은 종족 보존의 본능이다."에 대한 의존 문법 트리 구조와 구문 분석의 예를 보이고 있다.

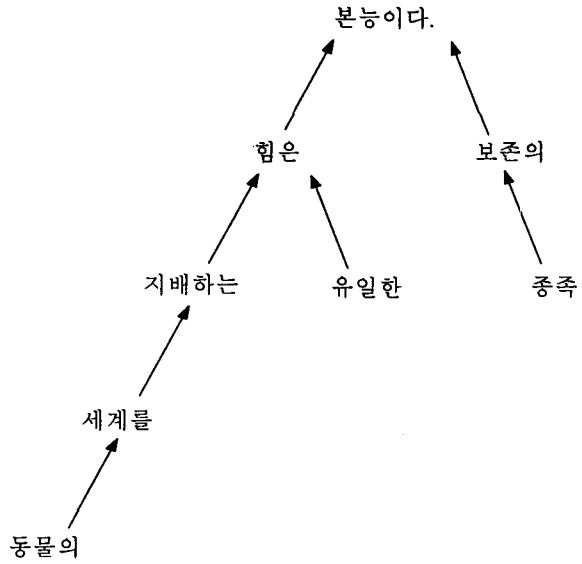

그림 1. 의존 문법에 의한 트리 구조
"동물의 세계를 지배하는 유일한 힘은 종족 보존의 본능이다."

표 72. 문장에 대한 구문 분석의 예

	Depth	RD	ToRight
동물의	4	0	0
세계를	3	0	0
지배하는	2	1	1
유일한	2	0	0
힘은	1	2	2
종족	2	0	0
보존의	1	0	0
본능이다	0		

3.4 어절의 위치

일반적으로 화자의 발음 스타일 또는 발음 속도에 따라 음절 혹은 어절의 절대적 길이는 변화하게 된다. 따라서 발음 시간에 관한 상대적 길이 정보를 얻어낼 필요가 있다.

문장에서의 어절의 위치는 하나의 문장의 발음 시간을 1로 보았을 때, 현재 어절의 상대적인 위치를 수치화 한 것이다, 휴지기로부터의 위치는 3.1절에서 언급한 방법으로 운율 경계 강도를 결정했을 때, 강한 경계('2')를 휴지기로 간주하여 휴지기로부터의 길이를 1로 보고, 현재 어절의 상대적인 위치를 수치화하였다.

표 73. 문장 sp0009.dat에 대한 파라미터 분석의 예
"동물의 세계를 지배하는 유일한 힘은 종족 보존의 본능이다."

어절	LPOS	POS	RPOS	RD	ToRight	BI	LOC	LOC_SP
동물의	OO	NG	NO	0	0	0	0.13	0.20
세계를	NG	NO	VA	0	0	0	0.25	0.40
지배하는	NO	VA	NG	1	1	1	0.38	0.60
유일한	VA	NG	NS	0	0	0	0.50	0.80
힘은	NG	NS	NG	2	2	2	0.63	1.00
종족	NS	NG	NG	0	0	1	0.75	0.33
보존의	NG	NG	VD	0	0	1	0.88	0.67
본능이다.	NG	VD	OO	0	0	0	1.00	1.00

IV. 억양의 성분 분리

음성 신호와 동시에 녹음된 Lx 신호의 파형을 자체 개발한 툴을 사용하여 분석한 후 각 문장에 대한 피치 정보를 별도의 파일로 기록하였다. 억양의 발생 모델을 3 단계, 즉 기저선 결정 단계, 어절의 평균 톤 결정 단계, 음절의 평균 톤 및 톤의 상승하강 결정 단계로 구분짓기 위해서는 구해진 피치를 3개의 주파수 영역으로 분리하여야 하는데 이를 위해 다음과 같은 한국어의 특징을 사용한다.

1. 음절의 평균 길이는 150~200 ms 정도이다.
2. 어절은 평균 3~4개의 음절로 구성된다.
3. 피치 신호는 500 Hz 이상의 고주파 성분을 갖지 않는 것으로 가정한다.

따라서, 음절톤은 기본 주파수가 5~6.7 Hz정도가 될 것이고 어절은 평균길이가 450~800 msec상에 존재하게 되므로 그 기본 주파수는 1.25 Hz~2.2 Hz정도가 될 것이다. 또한 음절톤을 기본 주파수의 5배수 정도까지 고려한다면, 약 30 Hz정도까지만 고려한다.

이런 가정 하에서 억양의 3성분을 분리하기 위한 필터링으로 Hanning 윈도우를 사용하는데, 이때의 천이구간(대역폭)은 다음의 식 (1)과 같다.

$$BW(2\omega_1) = \frac{8\pi}{M}, \quad M\text{은 필터의 차수} \tag{1}$$

한편, 피치를 10 msec마다 샘플링 한다고 할 때, 디지털 주파수 π는 50 Hz에 해당하므로 각 필터링을 위한 Hanning 윈도우 차수는 다음의 표 6과 같다.

표 74. 각 영역의 필터링 차수의 최적화

억양의 성분	Hanning 윈도우 차수
기저선(F_0_l)	8 π/M = (1.25~2.2) π/50 M = 181~320 (M = 301 적용)
어절 평균톤(F_0_m)	8 π/M = (5~6.7) π/50 M=66~80 (M = 80 적용)
음절 톤(F_0_h)	8 π/M = 30 π/50 M=13 (M = 13 적용)

다음의 그림 2는 문장에 대한 억양 성분 분리의 예이다.

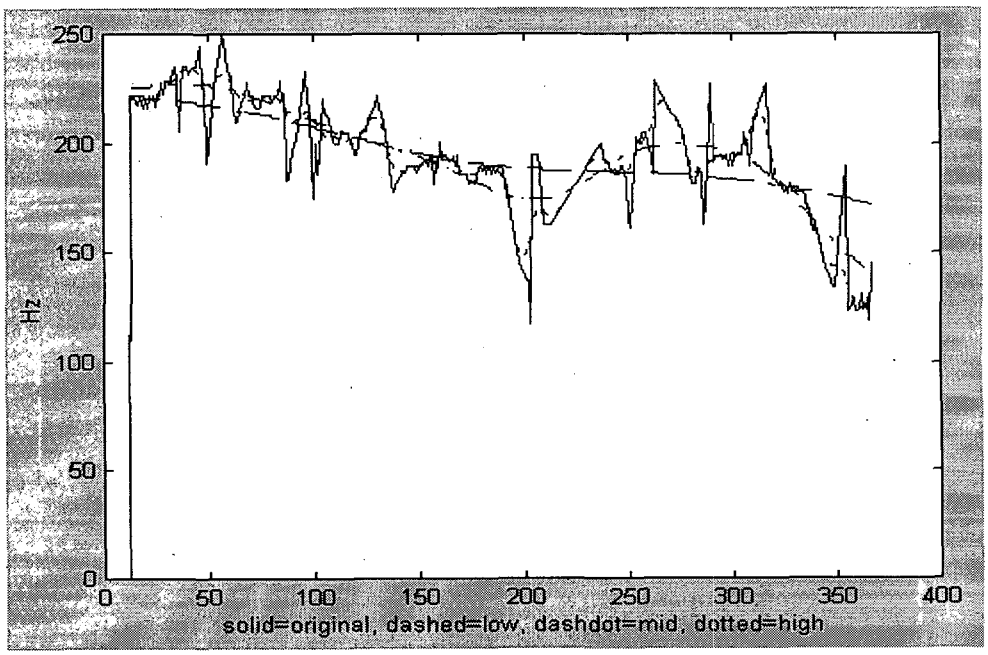

그림 103. 문장 spe0009.dat에 대한 억양의 성분 분리
"동물의 세계를 지배하는 유일한 힘은 종족 보존의 본능이다."

V. 어절 톤 예측 단계

본 논문에서는 어절톤을 예측하기 위해 CART(Classification And Regression Tree)를 사용하였다. CART는 크기의 순서가 정해져 있는 ordered 변수 이외에도 categorical 변수

에 대해서도 패턴인식 및 회귀분석을 적절히 수행할 수 있다는 장점을 가지고 있다. 본 논문에서 사용한 트리 기반 모델링 방법은 Gini 인덱스와 평균 제곱 오류를 이용하여 노드의 불순도(impurity)를 측정하고[25], Chou의 알고리즘에 의해 노드를 분할한 후[26], cost-complexity pruning 방법에 의해 트리를 선택한다[25].

어절 톤(F_0_m)을 예측하기 위한 방법으로 다음의 2가지 방법을 사용하여 이에 대한 성능을 비교하였다.

- 1-step method: 직접 어절 톤에 해당하는 피치(F_0_m)를 예측하는 방법.
- 2-step method: 기저선에 대한 어절톤의 비(F_0_m/F_0_l)를 예측한 후 기저선 피치(F_0_l)의 예측 결과를 곱하여 어절 톤의 피치를 예측하는 방법.
  즉, F_0_m = F_0_l*(F_0_m/F_0_l)

이때 각 피치 성분에 대한 예측 실험에서의 결과는 다음의 표 7과 같다. 그림 3은 한 문장에 대해 2가지 실험방법을 통해 예측한 어절 톤 결과를 나타내고 있다.

표 75. 어절 톤의 예측 오차

	1-step method F_0_m	2-step method F_0_m=F_0_l*(F_0_m/F_0_l)
평균 오차 (F_0_l, F_0_m/F_0_l)	12.37 Hz	12.41 Hz (9.82 Hz, 0.03)
최적 노드 수	47	49, 33

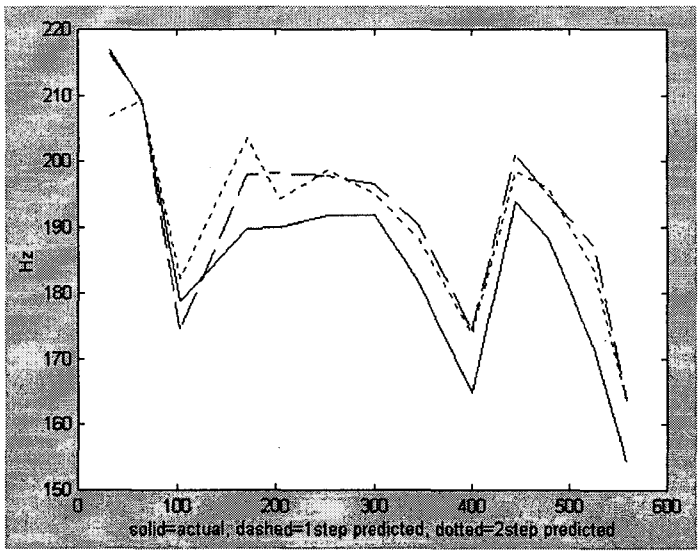

그림 104. 문장 spe1651.dat에 대한 어절 톤 예측 결과

"누구나 그렇게 살아 어느 누구도 자기의 삶의 주인공 자리를 다른 사람에게 내주진 않아."

표 7에서 보듯이 어절 톤을 예측하는 단계에서는 직접 어절 톤을 예측하는 경우(1-step method)와 두 단계로 나누어 예측하는 경우(2-step method)에 별다른 성능 차이를 보이지 않았다.

이때의 각 피치 성분의 예측 시 변수 중요도는 다음의 표 8~10와 같다.

표 8~10에서 볼 수 있듯이, 어절톤을 예측하는 단계에서는 상대적으로 어절의 위치와 품사정보가 중요한 파라미터로 작용함을 알 수 있다. 반면에 RD나 ToRight같은 구문구조와 관련된 파라미터는 상대적으로 중요도가 낮은 것으로 나타났다.

표 76. F_0_l 변수 중요도

변수	중요도
LOC	100.000
LOC_SP	39.264
POS2	21.895
LPOS2	14.600
BI	12.710
RPOS2	12.082
RPOS1	2.159
POS1	1.775
LPOS1	0.386

표 77. F_0_m/F_0_l 변수 중요도

변수	중요도
LOC_SP	100.000
POS2	52.351
LOC	47.024
BI	46.379
TORIGHT	18.870
LPOS2	12.822
RPOS2	3.451
POS1	1.226
LPOS1	0.458
RPOS1	0.089

표 78. F_0_m 변수 중요도

변수	중요도
LOC	100.000
LOC_SP	92.039
POS2	68.063
BI	28.827
LPOS2	22.364
RPOS2	12.698
POS1	5.940
TORIGHT	5.425
RPOS1	0.440
LPOS1	0.371

VI. 음절 톤 예측 단계

음절 톤(F_0_h)의 구현을 위해 본 절에서는 어절의 평균 톤(F_0_m)에 대한 음절 톤의 상승/하강 정도를 나타내는 차분값(F_0_h-F_0_m)을 사전을 구축하여 예측하였다.

억양을 구현하기 위해 운율구조를 표현하는 방안으로 ToBI 레이블링과 같은 방법을 사용할 수도 있으나[27-28], 악센트나 억양이 두드러진 특성을 지닌 영어나 일본어의 경우와는 달리, 한국어는 특징적인 어절 톤을 찾아보기 어려운 특성을 갖고 있으므로 본 논문에서는 음운학적 정보와 구문 분석 정보를 포함하는 사전을 이용하여 억양을 구현하고자 한다.

예측된 차분값을 앞서 예측한 어절의 평균 톤에 얹는 방법(F_0_h=(F_0_h-F_0_m)+F_0_m)으로 최종적인 음절 톤을 구현한다.

6.1 사전의 구축

본 논문에서는 음절 톤의 상승/하강 정도(F_0_h-F_0_m)를 예측하기 위하여 먼저 음절의 개수가 같은 어절들을 그룹화하여 사전을 구축하였다. 사전에 포함된 특징 파라미터는 해당 어절의 품사 정보와 구문 분석 정보, 운율 경계 강도, 어절의 시작점과 끝점의 피치값, 음운 환경 정보 등이다.

우리나라의 음절은 초성, 중성 그리고 종성으로 표현되므로 음운학적 정보는 한 음절에 대하여 세 개의 성분으로 나뉘어 구성된다. 또한, 초성을 예를 들어, 국어의 자소 자체를 특징으로 하여 어절톤 사전을 구축할 수도 있으나, 이러한 경우에는 자소 자체의 거리를 측정하는 것이 매우 어렵다. 따라서, 자소 자체를 해당 숫자로 고정시키는 방법(한글 조합형 표현과 같은)을 사용하지 않고 자소를 음운학적 특징으로 분류하는 방법을 사용하였다. 자음의 경우는 조음장소, 조음방법, 세기에 따라, 모음의 경우는 조음의 앞/ 뒤, 고/ 저, 원순, 이중모음의 여부에 따라 음운학적 특징을 분류하였다[29].

표 79. 3음절 단어의 사전 구조 1

단어	LPOS	POS	RPOS	BI	RD	ToRight	단어시작 F_0_h-F_0_m	단어끝 F_0_h-F_0_m	음절의 평균톤		
근래에	0	1	1	1	1	4	-6	10	-7.200	-10.727	-11.591
닭다리	6	9	1	0	0	0	3	2	-5.941	5.400	-0.818
요리한	4	9	14	0	0	0	-7	-2	-6.000	13.600	10.111
것이다	9	14	0	0	0	0	-2	-9	-2.111	-7.667	-10.538
원칙에	9	1	4	0	0	0	-11	2	-11.625	5.400	-0.308
제대로	4	5	1	0	1	1	5	-3	12.000	0.889	1.667
짓자면	1	4	9	2	3	4	-3	-15	-6.077	-0.100	-6.630
것이다	10	14	0	0	0	0	-9	-39	0.286	29.333	-38.471
음식의	0	9	1	0	0	0	-17	5	-15.286	14.917	8.583
재료와	1	1	9	1	2	1	-6	-4	1.429	2.053	-8.125
조리법	1	9	1	0	0	0	-4	-7	-3.143	12.353	1.846
:	:	:	:	:	:	:	:	:	:	:	:

표 80. 3음절 단어의 사전 구조 2

단어	음운 환경 정보		
근래에	{1,1,1}{3,3,1,1}{2,3,1}	{2,3,1}{4,1,1,1}{0,0,0}	{0,0,0}{4,2,1,1}{0,0,0}
닭다리	{2,1,1}{3,1,1,1}{1,1,1}	{2,1,3}{3,1,1,1}{0,0,0}	{2,3,1}{4,3,1,1}{0,0,0}
요리한	{0,0,0}{1,2,2,2}{0,0,0}	{2,3,1}{4,3,1,1}{0,0,0}	{5,6,1}{3,1,1,1}{2,2,1}
것이다	{1,1,1}{3,2,1,1}{0,0,0}	{4,5,1}{4,3,1,1}{0,0,0}	{2,1,1}{3,1,1,1}{0,0,0}
원칙에	{0,0,0}{3,2,2,3}{2,2,1}	{4,5,2}{4,3,1,1}{0,0,0}	{1,1,1}{4,2,1,1}{0,0,0}
제대로	{4,5,1}{4,2,1,1}{0,0,0}	{2,1,1}{4,1,1,1}{0,0,0}	{2,3,1}{1,2,2,1}{0,0,0}
짓자면	{4,5,1}{4,3,1,1}{2,1,1}	{4,5,3}{3,1,1,1}{0,0,0}	{3,2,1}{3,2,1,2}{2,2,1}
것이다	{1,1,1}{3,2,1,1}{0,0,0}	{4,5,1}{4,3,1,1}{0,0,0}	{2,1,1}{3,1,1,1}{0,0,0}
음식의	{0,0,0}{3,3,1,1}{3,2,1}	{4,5,1}{4,3,1,1}{0,0,0}	{1,1,1}{3,3,1,1}{0,0,0}
재료와	{4,5,1}{4,1,1,1}{0,0,0}	{2,3,1}{1,2,2,2}{0,0,0}	{0,0,0}{3,1,2,3}{0,0,0}
조리법	{4,5,1}{1,2,2,1}{0,0,0}	{2,3,1}{4,3,1,1}{0,0,0}	{3,4,1}{3,2,1,1}{3,4,1}
:		:	

※ {초성}{중성}{종성}의 순서

- 자음 : {조음장소, 조음방법, 세기}
- 모음 : {앞/뒤, 고/저, 원순, 이중모음}

6.2 음절톤의 상승/하강(F_0_h-F_0_m)

문장이 주어지면 그 문장의 각 어절에 대해 음절 톤 사전으로부터 다음 식 (2)의 거리 함수를 적용하여 최소 거리를 가지는 상위 N개의 후보 단어를 선정한다.

$$D_{total} = w_1 d_{phonenv} + w_2 d_{POS} + w_3 d_{BI} + w_4 d_{RD} + w_5 d_{toright} \quad (2)$$

$d_{phonenv}$: 음운 환경 거리 함수

d_{POS} : 품사 거리 함수

d_{BI} : 운율 경계 강도 거리 함수

d_{RD} : RD 거리 함수

$d_{toright}$: ToRight 거리 함수

이때 w_n은 각각의 파라미터의 거리 함수에 대한 가중치이며 그 최적값은 수 차례의 실험을 통해 얻어진 값으로, w_1=0.4, w_2=0.2, w_3=0.2, w_4=0.1, w_5=0.1이다.

위의 거리 함수 계산에 의해 선택된 최소 거리를 갖는 상위 N개의 단어를 대상으로, 운율 구 범위 내에서의 단어의 자연스러운 연결(smoothing)을 위해 단어의 경계에서의 피치간의 거리 함수를 최소로 하는 단어의 선택을 위한 Viterbi 탐색 알고리즘[30]을 사용하였다. 이때 단어 경계 간 피치 거리 함수는 다음과 같다.

$$D_{bound} = |F_0(last)_{i-1} - F_0(first)_i|^2 \quad (3)$$

$F_0(last)_{i-1}$: 앞 어절의 마지막 피치 차분값(F_0_h-F_0_m)

$F_0(first)_i$: 관측 어절의 첫 번째 피치 차분값(F_0_h-F_0_m)

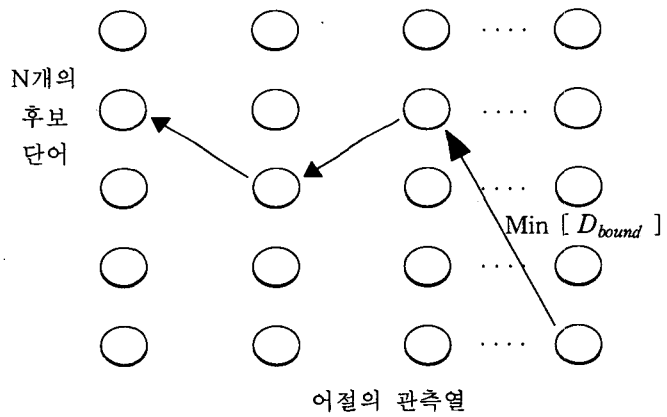

그림 4. 단어 경계 간 피치 거리 함수에 대한 Viterbi 탐색

실험 결과, 예측된 차분값(F_0_h-F_0_m)은 평균 8.66 Hz의 오차를 보였다.

6.3 최종적인 음절톤 예측

5절의 실험을 통해 예측된 어절의 평균 톤에 6.2.절에서 예측된 차분값(F_0_h-F_0_m)을 더한 최종적인 음절톤의 예측 결과는 다음의 표 13과 같다.

표 81. 음절 톤의 예측 오차

	1-step method F_0_h=(F_0_h-F_0_m)+F_0_m	2-step method F_0_h=(F_0_h-F_0_m)+F_0_l*(F_0_m/F_0_l)
평균 오차 (RMSE)	16.13 Hz (20.57 Hz)	16.08 Hz (20.52 Hz)

이때 여성화자의 평균 피치는 193.56 Hz이며, 평균 오차율은 약 8.3% 이내의 범위이다.

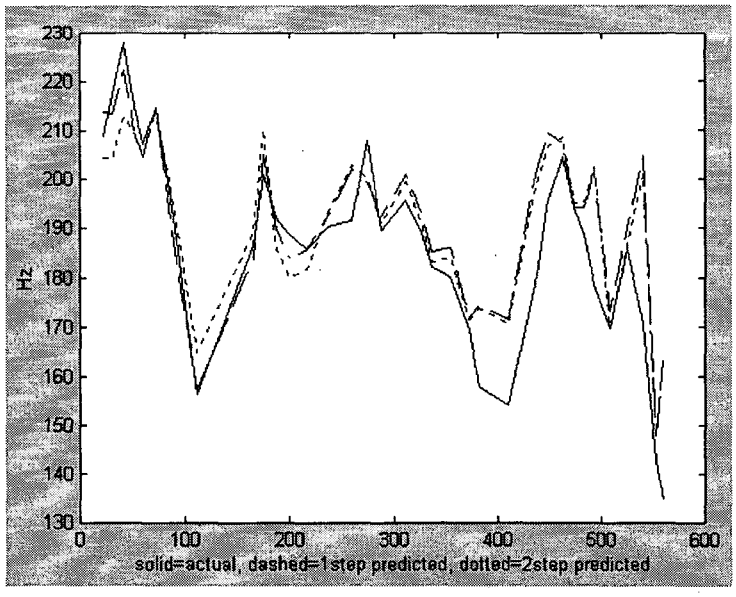

그림 106. 문장 spe1651.dat에 대한 음절 톤 예측 결과
"누구나 그렇게 살아 어느 누구도 자기의 삶의 주인공 자리를 다른 사람에게 내주진 않아."

VII. 결 론

본 논문에서는 자연스러운 합성음의 구현을 위해 억양성분을 3개의 피치 영역으로 분

리한 후, 어절 톤은 CART를 이용하여 학습한 후 2가지 방안으로 구현·비교하였으며 음절 톤에 대해서는 음절수에 따른 사전을 구축하여 단어 경계간 거리 함수에 대한 Viterbi 탐색 알고리즘을 수행하였다. 그 결과 어절톤의 예측 시에는 기저선과 기저선에 대한 어절톤의 상승 하강 정도를 구분하여 예측하는 2단계 작업은 직접 어절톤을 예측하는 작업과 비교하여 성능에 크게 차이가 없음을 보였다. 이때 변수 중요도를 살펴보면 해당 어절의 문장 내의 위치, 그리고 품사 정보가 중요한 파라미터로 작용함을 알 수 있었다. 최종적으로 구현된 음절 톤은 평균 오차율 8.3% 범위로서 비교적 양호한 성능을 보였다. 향후 본 논문에서 제시한 억양 구현 방안을 토대로 tri-phone 단위의 합성기를 제작하고 청취 실험을 통해 자연성 정도에 대한 평가를 실시하여 연구를 보완할 계획이다.

참 고 문 헌

[1] ETRI 보고서. 1995. 한국어 음성합성 기술 연구(V). 한국전자통신연구소.

[2] 이정철. 김상훈. 한민수. 1996. "한국어 대화체 음성합성기(글소리) 구현." *제9회 신호처리합동학술대회*. 49-52.

[3] 안승권. 1992. *한국어 음성변환 시스템에서의 합성음의 자연도 향상기법에 관한 연구*. 서울대학교 박사학위 논문.

[4] 구준모 외 5인. 1992. "한국어 무제한 음성합성 시스템: 가라사대." *제9회 음성통신 및 신호처리 워크샵 논문집*. 201.

[5] 김상룡. 김정수. 1993. "형태소 해석을 이용한 합성음성의 음운 및 운율처리." *전자공학회지 20-5*. 508-514.

[6] J. Pierrehumbert. 1981. "Synthesizing intonation." *J.A.S.A Vol. 70-4*. 985-995.

[7] M.D. Anderson, J.B. Pierrehumbert and M.Y. Liberman. 1984. "Synthesis by rule of English intonation patterns." *Proceedings of ICASSP 84.*. 2,8,1-4.

[8] K.N. Ross. 1986. *Modeling of Intonation For Speech Synthesis*. Ph.D thesis. Polytechnic University.

[9] P.C. Bagdshaw. 1994. *Automatic Prosodic Analysis for Computer Aided Pronunciation Teaching*. Ph.D thesis. Edinburgh University.

[10] A.J. Hunt. 1995. *Models of Prosody and Syntax and their Application to Automatic Speech Recognition*. Ph.D thesis. Sydney University.

[11] N.M. Veilleux. 1994. *Computational Models of the Prosody/Syntax Mapping for Spoken Language System*. Ph.D thesis. Boston University.

[12] K. Hirose, H. Fujisaki and M. Yamaguchi. 1984. "Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information." Proceeding of ICASSP 84. 2,13,1-2,13,4.

[13] H. Fujisaki, K Hirose and N. Takahashi. 1990. "Manifestation of linguistic and para-linguistic information in the voice fundamental frequency contours of spoken Japanese." Proceedings of ICASSP 90. 12,1-4.

[14] J. Allen, M.S. Hunnicutt and D. Klatt. 1987. *From Text to Speech: the MITalk System*. Cambridge University Press.

[15] D.B. Roe and J.G. Wilpon. 1994. *Voice Communication between Humans and*

Machines. National Academy Press.
[16] P. Taylor and S. Isart. 1994. "The rise/fall/connection model of intonation." Speech Communication. 169-186.
[17] 김진영. 성굉모. 1991. "한국어의 억양에 관한 연구." *Proceedings of Korea-Japan Joint Symposium on Acoustics*. 292-297.
[18] 이윤근. 안승권. 1993. "음성합성 기술 분야." *전자공학회지 20-5*. 523-532.
[19] 이정철. 김상훈. "최소 자승오차 방식을 이용한 세그먼트 피치패턴의 정형화." *제11회 음성통신 및 신호처리 워크샵 논문집*. 107-110.
[20] 김정수. 이혜정. 1996. "언어정보 및 통계데이터를 이용한 한국어 운율생성." *제13회 음성통신 및 신호처리 워크샵 논문집*. 227-231.
[21] 최운천 외 4인. 1992. "고품질의 한국어 문장음성변화 시스템: 글소리 II." *제9회 음성통신 및 신호처리 워크샵 논문집*. 193-196.
[22] Eric Sanders and Paul Taylor. 1995. "Using Statical Models to Predict Phrase Boundaries for speech synthesis." *Proceeding of EUROSPEECH 95 Spain*. 1811-1814.
[23] 김선미. 1997. *한국어의 리듬 단위와 문법 구조: 음성합성에서 리듬 구현의 자연성 향상을 위한 음성·언어학적 연구*. 서울대학교 박사 학위 논문.
[24] 김창현. 김재훈. 서정연. 1993. "지배 가능 경로를 이용한 오른쪽 우선 구문 분석." *제5회 한글 및 한국어 정보처리 학술 발표 논문집*. 35-44.
[25] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees*. Wadsworth Statistics/Probability Series, Belmont, CA.
[26] P.A. Chou. 1994. "Optimal Partitioning for Classification and Regression Trees." *IEEE Trans on PAMI Vol 13-4*. 304-354.
[27] Alan W. Black, Andrew J. Hunt. 1996. "Generating F0 Contours from ToBI lables using linear regression." *ICSLP Vol 3*.
[28] 이용주 외 6인. 1999. "K-ToBI 기호에 준한 F0 곡선 생성 알고리듬." *말소리, 35-36호*.
[29] 한국통신 연구개발본부. 1997. *음성 언어 시스템 개발을 위한 한국어의 운율구조 및 담화구조 연구*.
[30] Lawrence Rabiner, Biing-Hwang Juang. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall Inc. New Jersey. 339-340.

접수일자: 2000. 5. 10.
게재결정: 2000. 6. 04.

▲ 김 진 영
광주광역시 북구 용봉동 300
전남대학교 전자공학과(우: 500-757)
Tel: (062)530-1757 (O), Fax: (062)530-0472
e-mail : kimjin@dsp.chonnam.ac.kr

▲ 엄 기 완
광주광역시 북구 용봉동 300
전남대학교 전자공학과(우: 500-757)

Tel: (062)530-0472 (O), Fax: (062)530-0472
e-mail: eom@dsp.chonnam.ac.kr

▲ 박 상 언
광주광역시 북구 용봉동 300
전남대학교 전자공학과(우: 500-757)
Tel: (062) 530-0472 (O), Fax: (062) 530-0472
e-mail: separk@dsp.chonnam.ac.kr

▲ 최 승 호
전라남도 나주시 대호동 252
동신대학교 정보통신공학과(우: 520-714)
Tel: (0613) 330-3194 (O), Fax: (0613) 330-2909
e-mail: shchoi@white.dongshinu.ac.kr