피치 검색에 의한 Phrase 단위의 Template 추출에 관한 연구

김 종 국, 배 명 진
숭실대학교 정보통신공학과

On a Template Extraction of phrase unit by Pitch Searching

JongKuk Kim, MyungJin Bae
Dept. of Information&Telecommunication Eengr., SoongSil University
kokjk91@ssu.ac.kr

요약

원화자로부터 목표 화자의 음성으로 변환을 위해서는 음운 및 피치변환이 이루어져야 한다. 원 음성과 목표 음성 신호 사이에 따른 발성길이, 크기 및 피치 등의 운율 특성은 화자의 개인성 및 발성문장의 의도를 나타내는 주요 역할을 한다. 본 논문에서는 음성 변환을 수행하기 위하여 발성된 음성의 강세구(phrase)단위의 피치 검출을 통하여 템플릿을 추출하는 방법을 제안한다. 우선 한국어의 운율구에 대한 정보가 필요한 것인지, 한국어는 어떤 운율 구조를 갖는지에 대하여 알아본다. 마지막으로 어떻게 연속음성으로부터 한국어에 적당한 운율구 단위를 나눌 것인지, 즉 자동 세그멘테이션 및 레이블링에 대하여 분석한다. 또한 논문에서는 한국어 문장음성의 운율구를 강세구와 억양구로 나누고 육안으로 표시한 운율구 단위를 기준으로 이 운율구 단위에 적합한 특징을 추출하여 패턴을 작성한다

Ⅰ. 서론

일반적으로 글을 읽을 때나 자연스러운 대화를 할 때의 음성에는 각 화자마다 서로 다른 특성의 운율정보를 포함하고 있다. 운율정보는 단음절로 구성된 단어음성으로부터 시작하여 구나 절 단위의 음성에 이르기까지 발성된 음성신호에 포함되는 언어적 정보이면서 화자고유의 특성이기도 하다. 대화를 하거나 낭독을 하는 화자는 이 운율을 이용하여 의미, 감정, 의도 및 마음가짐 등을 전하며, 청취자는 화자가 발성한 음성으로부터 운율을 이용해서 전체적인 의미를 파악한다[1,2].

따라서 문장을 합성하는 경우 운율정보를 합성음에 반영하면 보다 명확한 의미 전달이 가능해 지며 화자의 음색을 잘 나타낼 수 있다. 자연 발화된 음성의 운율구 단위의 세그멘트와 레이블링은 연속음성의 인식에서 뿐만 아니라 음성합성 등의 음성신호처리 분야에서 필요한 처리 과정이다[1-3]. 왜냐하면, 인간은 음성언어의 경험적 지식을 바탕으로 음성의 음향학적 정보인 단시간 스펙트럼의 정보를 청각을 통하여 인식할 뿐만 아니라 악센트, 억양, 에너지의 크기 및 pause 등의 운율정보를 인지하여 음성을 인식하고 이해하기 때문이다.

언어학 분야에서는 운율 구조와 문장구조 및 음운규칙과의 관련성에 대한 많은 연구가 이루어져 왔다. 언어 이해 차원에서 의미 정보, 문장 구조 정보, discourse structure 등을 위한 운율 정보의 유용성이 입증되었으나, 이러한 결과가 최근의 음성인식 시스템에는 거의 적용되지 못하고 있다. 90년대 이후 미국의Ostendorf[4,5], 독일의 프로젝트인 Verbmobil[6] 및 일본의 Shimodaira[7,8] 등의 연구결과로 운율구 단위의 경계점 검출 및 레이블링 방법이 제시되면서 연속음성의 인식과 이해를 위한 작업으로 운율구 단위의 세그멘트 방법의 많은 연구가 시도되고 있다.

한국의 경우 운율 정보를 언어이해 시스템에 적용할 수 있는 가능성에 대한 연구가 계속적으로 이루어져, 한국어 중의성 문장의 음성을 이해하기 위하여 운율정보를 이용하여 의미를 추정하였다. 이외에 한국어 음성인식시스템에 운율정보를 이용한 예는 전무한 실정이다. 본 논문의 목적은 한국어의 연속음성인식 및 이해를 위한 실제적인 방법으로 한국어 운율구 단위를 자동 세그멘트하고 레이블링하는 데 있어서 다루어야 할 피치 추출 방법과 구(phrase) 단위의 템플릿 추출 방법에 대하여 제안하고자 한다.

본 논문에서는 한국어 문장음성의 운율구를 강세구와 억양구로 나누고 육안으로 표시한 운율구 단위를 기준으로 이 운율구 단위에 적합한 특징을 추출하여 패턴을 작성한다. 이 패턴을 이용하여 입력된 음성을 운율구 단위로 자동 세그멘트한다.

Ⅱ. 운율구

Ⅱ-1. 운율구의 필요성

하나의 단순문장에서도 휴지기의 위치에 따라 전체적인 의미가 달라짐을 알 수 있으며 또한 그에 따른 지속시간과 크기도 변화되는 것을 알 수 있다. 이러한 휴지기, 지속시간, 크기의 변화는 결국 문법적 구조와 관계되는 운율구의 정보를 제공하는 것이며, 청취자는 이 운율정보에 의해 운율구를 나누고 그에 따라 문장의 의미를 이해하게 되는 것이다. 그러므로 억양이나 휴지기, 지속시간, 크기와 같은 운율정보에 의해 구분되는 운율구는 연속음성인식을 위해 매우 중요하고 필수적인 요소임을 알 수 있다.

II-2. 한국어의 운율구조

사람이 사용하는 언어에서 문장보다 더 작은 단위로 생각할 수 있는 것은 우선 문법적 구 단위와 단어가 있다. 이것은 문장을 이루는 기본적인 단위로 이들에 의해 전체 문장의 구조가 형성된다. 문자 기반의 자연어처리 분야에서는 이러한 문법적 단위를 기준으로 문장을 분석하여 문장의 전체 구조를 파악한다. 그러나 이러한 문장들이 소리로 발화될 때는 사람의 호흡과 자연스럽게 조화되도록 문법적 단위와는 약간의 차이가 나는 새로운 단위로 재조정(readjustment)된다. 따라서 음성 기반의 음성언어처리에서는 문법적 단위가 아닌 음향적 특성에 의해 나타나는 새로운 단위가 필요하다.

전 세계의 언어를 총괄적으로 생각해 보면 Nespor와 Vogel이 제안한 계층적 운율구조가 모두 나타나고 있으나, 각 국의 언어마다 이들의 계층구조가 그대로 적용되는 것은 아니다. 언어학자 전선아[9]는 이들의 계층구조가 7가지 단위로 구성되었지만 각 국의 언어에 모든 단위가 필요한 것은 아니며 이들 중 몇 가지만 있어도 각 국의 언어에 적당한 운율 구조를 구성할 수 있다고 말하고 있으며, 한국어의 경우 강세구(accentual phrase)와 억양구(intonational phrase)가 한국어의 기본적 운율 구조를 구성하고 있다고 제안했다.

본 논문에서는 운율 구조 이론을 받아들여 한국어 연속음성을 위한 운율구 단위의 자동 세그멘테이션과 레이블링 알고리즘의 언어학적 기반으로 삼았다. 다음의 그림1은 한국어 운율 구조를 보이고 있다.

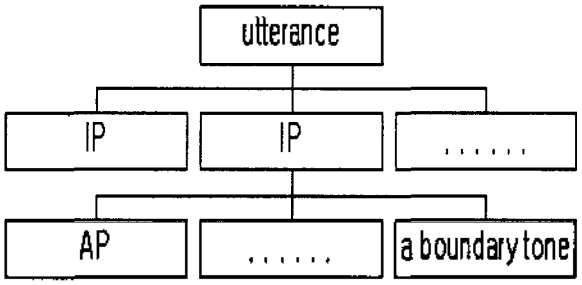

그림 1. 한국어 운율구조

이 그림의 억양구(intonational phrase)는 Nespor와 Vogel이 제안한 보편적 계층구조와 일치하는 개념으로서, 한국어의 경우 톤레벨(tone level)이 H(igh)와 L(ow)로 구성되는 억양궤적에 의해 정의된다. 여기서 억양궤적은 피치 엑센트(pitch accent)와 어절톤(phrase tone)으로 구성되며, 피치 엑센트란 어휘 고유의 엑센트(lexical stress)이며 어절톤은 어절의 엑센트와 경계톤(boundary tone)으로 구성된다. 어절의 엑센트는 억양구보다 작은 단위이며, 경계톤은 억양구 단위를 구분하는 부분이다. 이와 같이 운율구 단위를 억양구와 그보다 작은 단위로 나눌 수 있으며 억양구 단위보다 작은 단위를 강세구(accentual phrase)라 한다.

III. 운율구의 검출

본 논문에서는 한국어 운율구 단위를 자동 세그멘트하기 위한 기본 단위로 억양구와 강세구 단위를 사용한다. 억양구 단위의 세그멘테이션 알고리즘은 휴지기 또는 묵음구간의 검출에 의해 이루어지며 대개의 경우 문장경계나 억양구 단위의 경계를 세그멘트해준다. 강세구 단위의 세그멘테이션 알고리즘은 미리 수동적으로 세그멘트된 강세구 단위의 피치패턴과 DTW 알고리즘을 이용하여 억양구 안의 강세구 단위를 세그멘트한다. 즉, 음성이 입력되면, 우선 휴지기를 이용하여 입력음성을 억양구 단위로 세그멘트하고, 이 억양구 단위 음성의 피치패턴과 강세구 단위 피치패턴 사이의 세그멘트 경계를 검출한다.

III-1. 억양구(Intonational Phrase)

문장을 발성할 때 음성은 자연스럽게 휴지기를 갖으며 그에 의해 문장의 의미도 달라질 뿐만 아니라 이를 연속음성의 인식에 이용하면 인식성능과 계산량의 감소에 효과적인 역할을 한다. 본 연구에서는 억양구를 자동 세그멘트하기 위하여 에너지 특징을 이용하여 휴지기를 검출한다. 단, 음성의 묵음구간은 문장의 경계나 억양구 단위의 경계에 나타나는 휴지기이기도 하지만 파열음의 앞/뒤에 생성되는 요소이기도 하다. 따라서 억양구 단위의 자동 세그멘테이션을 위해서는 에너지 특징에 의한 묵음구간이 휴지기인지 아닌지를 검토해 줄 필요가 있다.

또한 억양구는 문장 중간에 나타나는 억양구와 문장말에 나타나는 억양구로 나눌 수 있다. 이들은 다시 의미에 따라 세분화된다. 문장말의 억양구는 의미에 따라 여러 가지 다른 피치궤적으로 나타난다. 예를 들어, 애매성 문장의 음성에서 문장말 억양구는 의미에 따라 피치궤적이 서로 다르게 나타나고 있다. 이러한 문장말 억양구의 특징은 애매성 문장이 아니더라도 서술형, 의문형, 명령형 및 권유형 문장의 연속음성에 모두 공통적으로 나타나기 때문에 문장말의 억양구는 각 의미에 따라 세분화가 가능하며 이에 따라 레이블링이 가능해진다.

III-2. 강세구(Accentual Phrase)

억양구를 형성하는 단위인 강세구는 주로 피치 궤적(f0 contour)에 의해 특징지어진다. 하나의 강세구가 세 음절 이하로 구성될 경우 주로 L H (low-high)의 피치 궤적으로 나타나며, 네 음절 이상일 경우 L H L H의 피치 궤적으로 나타난다. 하나의 억양구 내에서 여러 개의 강세구가 나타날 경우는 피치 궤적이 점차적으로 낮아지는 단계적 하향현상을 보인다. 또 억양구의 마지막 강세구에는 경계톤이 기본 강세구의 피치 궤적에 덮혀짐으로써 기본 L (H L) H 피치 궤적이 아닌 경계톤의 피치 궤적으로 나타나게 된다.

문법적 어절의 경계와 운율적 어절의 경계는 매우 밀접한 관계가 있다. 한국어 운율구조에서의 강세구 단위의 주요 특징은 피치궤적이므로 그림 2와 같이 강세구 단위로 hand labeling한 후 그의 피치궤적을 표준패턴으로 사용한다. 이러한 표준패턴을 기준으로 패턴처리 기법을 이용하여 강세구를 세그멘트한다[12].

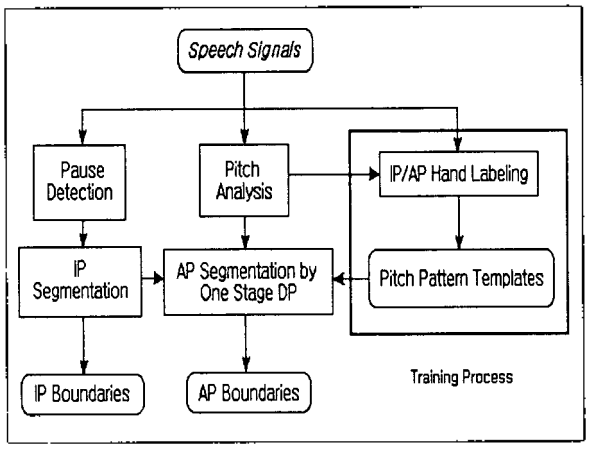

그림 4. 운율구 검출 시스템

III-3. 운율 특징의 분석

III-3-1. 피치 분석

하나의 호흡단위 내에서 발성된 문장 음성의 피치궤적은 시간이 지날수록 점점 낮아지며 굴곡의 폭도 차츰 좁아진다. 이는 생리적 현상 중에서 가장 뚜렷한 특징이다. 또한 한 문장의 연속음성은 여러 개의 강세구(accentual phrase)로 나뉘어 지기 때문에 피치궤적으로 나타나는 강세구의 특징을 보다 분명히 하기 위해서는 피치궤적의 정규화가 더욱 필요하다. 한 문장 음성 파형에 대한 피치궤적, 시간에 따른 피치의 최소값으로 구성된 기저선(baseline) 및 정규화된 피치궤적을 보이는 이 기저선은 거의 일 직선상에 존재하며 부 방향의 기울기를 가진다. 이 기저선은 발성자나 음운학적 조건에 따라 다를 수 있으나 근사적으로 다음 식과 같이 나타낼 수 있다[11].

$$B(t)=\frac{p(T)-p(0)}{T}t+p(0) \qquad (1)$$

여기서, $p(0)$ 은 음성의 시작점에서의 피치 주파수이고 $p(T)$은 음성이 끝나는 시간에서의 피치 주파수이며, T는 음성의 시작점에서 끝점까지의 총 시간길이 이다. 이상의 식(1)을 이용하여 제안한 정규화된 피치궤적을 식으로 나타내면 다음과 같다.

$$\rho(t)=p(t)-B(t)(0\le t\le T) \qquad (2)$$

III-3-2.. 패턴 매칭에 의한 강세구 검출

운율구 단위로 제안한 강세구(accentual phrase)를 사용하여, 문장 단위의 음성 데이터를 L (H L) H 패턴의 피치 궤적(pitch contour)에 의해 강세구로 나누어 경계점을 검출한다. 이 때 피치 궤적의 패턴 매칭 기법으로는 DP 알고리즘[12]을 사용한다. DP 알고리즘에 사용할 강세구의 표준 패턴과 시험 패턴을 작성하기 위하여 음성이 입력되면 먼저 피치를 추출하고 이를 정규화한다. 또한 L (H L) H 패턴의 피치 궤적을 수동으로 채취하여 표준패턴으로 하고 입력된 음성 신호의 정규화된 피치궤적을 시험패턴으로 한다. 연속음성 데이터에서 강세구의 경계점을 검출한 후에는, 그 경계점에서부터 역방향으로 추적하여 문법적 기능어를 인식할 수 있다.

IV. 실험 및 결과

IV-1. 실험 조건

본 실험에서는 표준어 말씨의 남자 10명과 여자 5명이 사전 지도 없이 서술형 문장 20개를 낭독하여 각자의 녹음기에 녹음한 것을 실험 음성 데이터로 하였다. 낭독할 서술형 문장 중에서 4문장은 4개의 강세구, 다른 4문장은 5개의 강세구, 나머지 12개 문장은 3개의 악센트구로 구성되어 20개의 문장에 포함된 악센트구의 수는 총 72개이다. 따라서 본 실험의 운율분석에서 세그멘트해야 할 악센트구의 수는 모두 1080개이다. 컴퓨터 저장을 위해 KAY사의 Multi-speech를 이용하여 10kHz로 샘플링 하였으며 분석을 위한 프레임 간격은 25.6msec, 이동간격은 12.8msec로 하였다.

운율 분석 단위인 강세구를 자동 세그멘트하기 위하여 one-stage DP를 이용하였다. 표준패턴은 남자 2명과 여자 1명이 각각 발성한 1개의 문장에서 수동으로 검출한 강세구를 multi-pattern으로 구성하였다. 운율분석은 그림 2의 강세구 세그멘트 시스템에 의해 실험되었다. 시스템에 음성이 입력되면, 우선, 피치를 추출하여 정규화하고 같은 방법으로 정규화된 강세구 단위의 표준 피치패턴 사이의 one-stage DP를 이용하여 입력 음성의 강세구를 자동 세그멘트한다.

IV-2. 패턴 매칭에 의한 강세구 검출 결과

연속음성이 입력되면 우선 피치를 분석하고 정규화된 피치 궤적과 one-stage DP를 이용하여 강세구의 경계점을 자동 검출한다. 여기서 사용한 음성 데이터는 72개의 문법적 어절을 포함하는 20개의 평서문 문장이며 표준 패턴은 남자 2명과 여자 1명이 각각 발성한 1개 문장에서 수동으로 검출하여 레이블링한 강세구의 정규화된 피치패턴으로 구성하였다.

다음의 그림 3과 4의 점선으로 연결된 세모 표시는 패턴 매칭을 사용하여 즉, one-stage DP를 사용하여 여자 5명과 남자 10명의 각 화자에 따른 강세구의 경계를 검출한 결과이다. 여성 화자 전체의 평균 검출률은 80.8%이

며 남성 화자 전체의 평균 검출률은 83.3%이다. 그림 3과 4의 직선으로 연결된 원형 표시는 LH톤을 탐색하여 각 화자에 따른 강세구의 경계를 검출한 결과이다. 여성 화자 전체의 평균 검출률은 74.2%이며 남성 화자 전체의 평균 검출률은 75.0%이다.

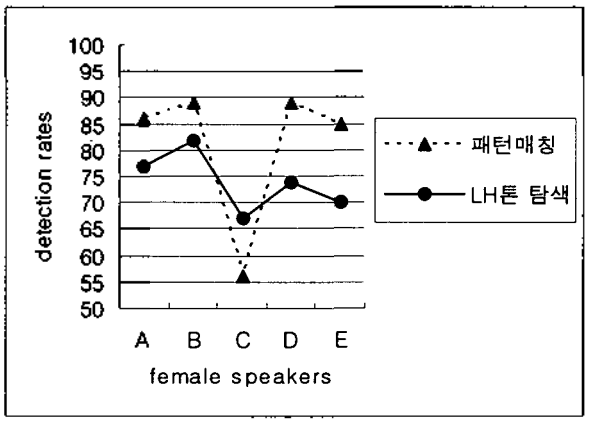

그림 5. 여성 화자별 검출률 비교

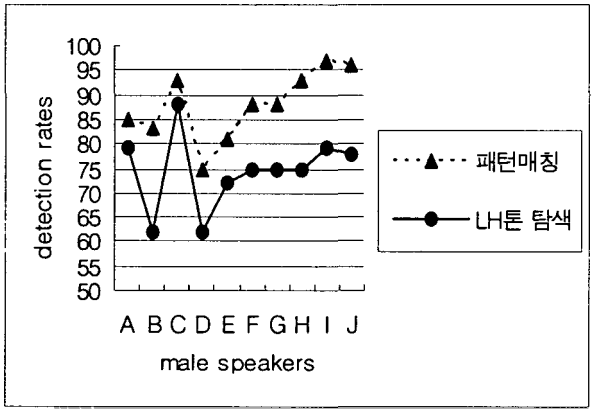

그림 6. 남성 화자별 검출률 비교

V. 결론

본 논문에서는 발성 변환을 위하여 연속음성의 운율분석을 통해 운율구 단위의 템플릿을 추출하는 방법에 대하여 제안하였다. 제안한 피치 검출방법을 이용하여 연속음성을 강세구 단위로 운율 분석하였다. 강세구의 검출률은 남성화자의 경우 약 88.3%, 여성화자의 경우 약 80.8%를 얻어 총 평균 85.8%의 검출률을 보였다. 따라서 본 논문에서 제안한 템플릿 추출 방법이 운율분석에 적당함을 확인할 수 있었다.

패턴 매칭에 의한 모든 경우에 있어서 강세구의 경계는 문법적 어절의 마지막 음절의 경계와 일치하지 않았다. 더구나 피치 궤적을 참조로 하여 검출한 강세구의 경계는 문법적 어절의 마지막 음절의 경계와 대부분 일치하지 않았으므로 음절 분석 결과로 나타나는 음절 경계를 참조하여 가까운 위치로 강세구의 위치를 이동해 주면 문법적 어절의 경계와 정렬시킬 수 있다. 이를 개선하기 위해서는 본 논문에서 제안한 운율구 단위의 템플릿 추출에 있어서 강세구의 자동 검출 방법과 보다 향상된 음절 경계의 추출 방법을 조합하여 강세구 경계의 정확성을 보완해 줄 필요가 있는 것으로 생각된다.

감사의 글

본 연구는 한국과학재단 특정기초연구(과제번호: R01-2002-000-00278-0)의 지원에 의하여 이루어졌음.

참고문헌

[1] E. O. Selkirk, *Phonology and Syntax*, The MIT Press, Cambridge, Massachusetts, 1984.

[2] M. Nespor and I. Vogel, *Prosodic Phonology*, Foris Publications, Dordrecht, 1986.

[3] M. Beckman and J. Pierrehumbert, "Intonational structure in Japanese and English," *Phonology Yearbook 3*, ed. J. Ohala, pp. 255-309, 1986.

[4] C. W. Wightman and M. Ostendorf, "Automatic labeling of prosodic patterns," *IEEE Trans. Speech, Audio Processing*, Vol. 2, No. 4, pp. 469-481, 1994.

[5] C. W. Wightman and M. Ostendorf, "Automatic recognition of intonational features," in *Proc. of IEEE Int. Conf. ASSP*, pp. 1-221, 1992.

[6] A. Kipp, M. Wesenick, and F. Schiel, "Automatic detection and segmentation of pronunciation variants in German speech corpora," in *Proc. of ICSLP96*, 1996.

[7] H. Shimodaira and M. Kimura, "Accent phrase segmentation using pitch pattern clustering," in *Proc. of IEEE Int. Conf. ASSP*, pp. I-217, 1992.

[8] H. Shimodaira and M. Nakai, "Prosodic phrase segmentation by pitch pattern clustering," in *Proc. of IEEE Int. Conf. ASSP*, pp. II-185, 1994.

[9] Sun-Ah Jun, *The Phonetics and Phonology of Korean Prosody*, Doctoral Dissertation, The Ohio State University, 1993.

[10] KiYoung Lee, MyungJin Bae, HoYoung Lee, JongKuk Kim, " Pitch Contour Conversion Using Slanted Gaussian Normalization Based on Accentual Phrases ", Korean Journal of Speech Sciences, Vol. 11, No 1, pp. 31-41, 2004.

[11] J.K. Kim, M.J Bae, " A study of Pitch Extraction Method by using Harmonics Peak-Fitting in Speech Spectrum ", in Proc. of ICSP2001 Conf. , Vol. 2, pp. 617-621,2001.

[12] JongKuk Kim, KiYoung Lee, MyungJin Bae, "On a Detection of Korean Prosody Phrases Boundarie", IEAL2004, LNCS3177, pp.241-246,2004.