문장 단위 운율 제어를 위한 신경망의 입력 패턴에 관한 연구

A Study on the Input Pattern of Neural Network for Prosody Control in a Korean Sentence

민경중* 임운천
호서 대학교 대학원 전자공학과

Kyoung-Joong Min*, Un-Cheon Lim
Department of Electronic Engineering, A Graduate School, Hoseo University
E-mail: uclim@dogsuri.hoseo.ac.kr

요약

법칙 합성 시스템은 합성 단위, 합성기, 합성방식등 여러 가지 다양한 시스템이 있으나 순수한 법칙 합성 시스템이 아니고 기본 합성 단위를 연결하여 합성음을 발생시키는 연결 합성 시스템은 연결 단위사이 그리고 문장 단위에서의 매끄러운 합성 계수의 변화를 구현하지 못해 자연감이 떨어지는 실정이다. 자연감에 영향을 끼치는 주요 원인중의 하나가 운율 법칙의 부정확한 구현이므로 자연음으로부터 추출한 운율에 관한 법칙을 알고리듬화하는 대신 신경망으로 하여금 이 운율 법칙을 학습하도록 하여 좀더 자연음의 운율에 근접한 운율을 발생시키고자 하였다. 신경망으로 운율을 발생시키기 위해 먼저 운율에 영향을 주는 요소들을 정해 신경망 입력 패턴을 선정해야 한다. 먼저 분절요인에 의한 영향을 고려해주기 위해 전후 3음소를 동시에 입력시키고 문장내에서의 구문론적인 영향을 고려해주기 위해 해당 음소의 문장내에서의 위치, 운율구에 관한 정보등을 신경망의 입력 패턴으로 구성하였다.

Ⅰ. 서 론

음성합성은 인간의 음향학적 정보 전달 수단인 음성을 기계가 소리의 합성을 통하여 발생시키는 기술이다. 이 기계에 의한 합성음은 올바른 정보 전달 능력으로서 이해도와 인간의 발성과의 유사함을 나타내는 자연성으로 평가되어 진다. 또한 음성합성의 영역이 넓어지고 보편화됨에 따라, 인간의 음성과 같이 자연스러운 합성음에 대한 요구가 증가되고 있다.

법칙 합성 시스템은 합성 단위, 합성기, 합성방식등 여러 가지 다양한 시스템이 있으나 순수한 법칙 합성 시스템이 아니고 기본 합성 단위를 연결하여 합성음을 발생시키는 연결 합성 시스템은 연결 단위사이 그리고 문장 단위에서의 매끄러운 합성 계수의 변화를 구현하지 못해 자연감이 떨어지는 실정이다. 특히 시간 영역 법칙 합성 시스템의 합성음은 이해도는 향상되었음에도 불구하고 자연감이 많이 떨어지고 있다[8].

문장내에서 적절한 운율구의 경계를 추출하면, 이러한 경계의 전후 음절의 피치변화, 앞음절의 장음화, 음질 변화등을 고려하여 자연감을 위해서 다시 변화 법칙을 추출하고, 이것을 다시 알고리듬으로 작성해서 재 평가하는 과정을 반복해야 한다. 이러한 복잡한 알고리듬화 대신, 문장의 마침표와 같은 확실한 경계가 있는 구간과 쉼표와 같은 약한 경계, 그리고 blank와 같은 어절간의 경계를 신경망에 학습시킴으로써 복잡한 운율법칙 알고리듬을 위한 프로그램 작성없이 연속음의 운율변화곡선을 발생시키는데 필요한 신경망의 입력패턴의 구성방법을 제안하고자 한다.

Ⅱ. 운율제어

일반적으로 서로간 자연스러운 대화라든가 글을 읽

을 때의 음성, 즉 자연음은 화자의 감정상태, 말의 내용 또는 강세, 발음속도와 같은 의미론적인 정보와 전체의 구문구조와 문장내에서의 구와 절의 경계위치, 기능, 상호 결합관계등의 구문론적인 정보가 있으며, 단어에서의 강세유형과 각 분절음소의 전후 결합에 의한 영향이 음향학적 특성을 결정하게 된다. 이와 같은 여러가지 요인에 의해 같은 음소일지라도 문장내에서의 위치에 의해 피치와 지속시간, 크기 등이 달라지는데 이들 피치와 지속시간 및 크기 등의 변화를 운율이라 한다.

의미론적인 정보에 의한 문장단위의 운율법칙을 추출하기 위해서는 여러 가지 상태에 따른 많은 양의 데이터와 오랜 처리시간이 필요하고 이러한 운율정보를 법칙화하는데 많은 어려움이 따르게 된다. 그러나 구문론적인 정보는 문장의 구조를 해석하여 구와 절 또는 운율구, 억양구 등의 경계를 분리해내어 구 단위의 운율을 생성하여, 문장단위에서 보다 쉽게 법칙을 추출하여 운율을 제어할 수가 있다.

Ⅱ-1. 운율의 3요소

1. 피치변화

피치가 음성의 자연감에 미치는 영향은 매우 큰 것으로 평가되고 있다. 피치의 변화에 영향을 주는 요인으로 먼저 의미론적인 요소 즉 대비, 강조, 화자의 감정상태와 발음속도등이 있으나 이들이 미치는 변화를 모델링 하기에는 너무 많은 데이터처리가 필요하기 때문에 일반적으로 피치모델을 작성할 때에는 중립상태에서 보통속도로 발음하는 것을 대상으로 하고 있다. 다음으로 구문론적인 측면에서는 일부 예외가 있으나 대부분 평서문을 대상으로 피치모델을 구하고 있어 실제 대화체의 자연음에서의 피치변화를 모델링할 수 있는 단계에는 아직 미치지 못하고, 평서문에서 구문의 구,절 등의 경계와 단어의 강세유형 그리고 분절음소가 미치는 영향을 고려한 피치모델을 구하고 있다.

피치변화 모델을 세분하면 문장 전체에 걸친 피치의 주 변화와 구.절 등의 경계와 단어의 강세유형에 따른 부 변화 그리고 분절에 의한 미세 변화로 나눌 수 있다[2].

2. 지속시간의 변화

지속시간을 변화시키는 요인으로 전후음소에 의한 영향, 강세의 유·무에 의한 영향, 휴지전 여부에 의한 영향, 음절수에 의한 영향, 단어의 빈도수에 의한 영향 등이 있다[2].

영어에서는 뒷자음이 무성 중지음일 때 지속시간이 가장 짧고 유성마찰음일 때 가장 길며, 강세모음과 비강세 모음의 지속시간의 차이는 구의 끝에서만 나타난다. 또한 휴지전 음절이 길어지는 휴지전 장음화 현상이 나타나고 음절수에 따른 변화는 음절수가 늘어남에 따라 줄어드나 선형으로 줄어들지는 않으며 빈도수가 낮은 단어일수록 길어지는 경향을 보이고 있고, 한 음절이 처음, 중간, 끝에 위치할 때 각각의 지속시간에는 차이가 있으며 일반적으로 음절수가 많은 단어의 끝음절은 다른 음절보다 지속시간이 긴 경향이 있다.

한국어의 지속시간에 대한 연구는 2음절로 구성된 무의미 단어에 대한 발음을 대상으로 하여 유성자음에 비해 무성자음 앞에 오는 단음의 지속시간이 짧아지고 마찰음에 비해 중지음 앞의 모음길이가 짧아짐을 보고하고 있고, 6음절 문장을 대상으로 하여 전체 문장 길이의 분산이 각 음절 길이의 분산의 합에 비해 적어 음절이 기본단위가 아니고 분절이 기본단위임을 제시하고 있다.

3. 에너지의 변화

음성 신호는 발성 모델의 여기음에 따라 성대진동의 기본주기인 피치주기와 성도 공명에 의한 큰 에너지 성분을 갖는 유성음과 성대 진동을 수반하지 않으면서 단지 성도의 조음점을 스치고 지나가기 때문에 그 파형의 에너지가 작은 무성음으로 크게 분류할 수 있다. 묵음 부분은 초성, 종성의 폐쇄음에서 나타낸다. 에너지는 음소를 구별하는데 중요한 운율요소 중에 하나이고, 합성음의 자연성에 미치는 영향이 큰 강세 및 억양등에 구성되는 운율적 요소로 합성음을 법칙화하는데 중요한 요소이다[5].

Ⅱ-2. 문장내에서의 운율변화

1. 악센트

악센트는 단어단위의 음성에서 앞뒤소리와 상대적인 관계에 의한 어느 한 음절을 강조함으로써 뜻의 전달에 기여하는데, 악센트는 피치, 지속시간, 세기에 의해 구성되어 진다. 일반적으로 악센트는 한 문장의 표면구조인 한 개의 음형 악센트로 되어 있고, 국어의 악센트특징은 첫음절의 강세와 높이이다. 즉, 낭독형은 중간높이의 약한음절로, 강조형은 중간높이의 강한음절로, 대조형은 높고 강한 음절로 시작하는 것이다. 강음절은 강조형과 대

조형의 첫음절에 나타날뿐 나머지는 모두 약음절이다. 또한 고음절은 낭독형의 중간부분, 강조형의 끝부분, 대조형의 첫부분에 나타난다. 일부 악센트는 의미적 변별 기능을 갖는데, 예를 들어 눈(eye)과 눈(snow)등이, 이에 속한다. 이러한 악센트는 합성음의 자연성에 영향을 주는 갖는 중요한 요인 중에 하나이다[8][11].

2. 억양

문장단위의 음성에서 피치가 움직이는 방향과 악센트를 받는 음절간의 상대적인 음높이와 같은 피치의 성질을 '억양'이라 정의하는데, 억양은 피치의 변화로만 구성되는 것이 아니라 지속시간, 강세, 리듬, 속도, 목소리의 음질 등의 음향적 요소가 결합하여 만들어진 것이라고 할 수 있다. 그러나, 일반적으로 억양은 주로 피치로 실현되고 경우에 따라 크기나 지속시간이 수반된다. 한국어에 있어서 문장의 억양은 의미를 변화시키지는 않지만 문장의 형태, 구문구조, 화제의 변경, 의미, 감정 등에 따라 다양한 형태로 나타난다. 문장 중간에서 실현되는 억양은 억양구내에서 그 구의 구문론적, 의미론적인 정보를 나타내어 준다. 예로 한국어의 문장끝의 억양구 단위에서 피치 곡선은 평서문, 의문문, 감탄문은 오르거나 내리거나 평탄하거나 하는 한 가지 방향만의 피치를 가지고 있다[6][10].

3. 운율구

화자의 자연스러운 발성에 따라 형성되는 단위로서 피치, 지속시간 ,크기, 억양 등의 운율변화로 나타나므로, 문장을 적절한 운율구로 나누어서 분석하므로써 음성합성의 자연성을 크게 향상시킬 수 있다. 문장내에서 운율구의 경계를 추출하여 구 단위로 분석하므로서 문장단위의 여러 가지의 운율 현상을 세분하여 합성음의 자연성의 규칙화를 쉽게 할 수 있게 된다.

운율구는 문장내에서 화자와 청취자가 모두 객관적으로 동의할 수 있는 음성학적 끊김이 있는 단위로서, 일정 수치 이상의 지속시간의 증가나 피치의 변동으로 경계지어진 단위이다. 즉 분명한 끊어 읽기가 이루어진 큰 음성학적 단위의 설정이 바로 운율구라고 하였다. 운율구 한 개에 포함되는 음절수는 그 빈도에 있어 대략 5-8음절 정도가 대종을 이룬다고 할 수 있다[1][3].

3-1. 경계의 정의

자연스러운 합성음을 생성해내기 위해서는 먼저 하나의 문장이 들어 왔을 때 이를 어떻게 몇 개의 경계로 나누어 줄 것인가 하는 것이 문제이다. 일반적인 경계는 단어경계와 구,절의 경계로 나눌 수 있지만, 인간이 말을 할 때는 언어외 통사 의미 구조에 관한 지식, 문장의 길이, 말의 속도, 심리적 생리적 요인 등 다양한 요인을 가지고 자연스러운 말의 패턴을 만들어내기 때문에, 이러한 작업을 기계에서 처리하기 위해서는 인간이 가지고 있는 문장단위에서의 음성 생성에 관한 지식을 명시적이고 구체적인 정보로서 제시할 필요가 있는데, 이러한 자연성을 가진 연속음의 경계를 어떻게 나누어서 분석하느냐에 따라 합성 법칙이 결정되는 것이다. 이러한 경계를 구하는 방법은, 크게는 문장내에서 화자와 청취자간 모두 객관적으로 동의 할 수 있는 음성적 끊김이 있는 단위로 경계지어진 운율구[3]와 breath group이 되는 구내에서 그 구의 억양에 의한 구문론적, 의미론적인 정보를 나타내어 주는 억양구[6]가 있다. 그 이외 좀더 세분화 시킨 알고리듬도 있으나, 이러한 것은 문법적인 구조 및 화자의 감정상태, 발음속도 등에 따른 변화가 많으므로, 본 논문은 문장의 마침표와 같은 확실한 경계가 있는 구간과 쉼표와 같은 약한 경계, 그리고 blank와 같은 어절구간의 경계를 나누어서, 이것도 하나의 음소화하여 신경망을 학습시킴으로써 복잡한 운율 법칙 알고리듬을 위한 프로그램 작성없이 연속음의 운율변화곡선을 발생시키는데 필요한 신경망의 입력패턴의 방법을 만드는데 목적이 있다.

Ⅲ. 신경망 구성

신경망 시스템은 계층적 구조인 입력층, 은닉층, 출력층의 3개의 층으로 구성하였다. 입력층은 7개의 음소열를 나타내는 유니트들로 이루어져 있는데 각 음소는 16개의 유니트로 구성하고, 이를 위해 얻어진 음소 즉 초성 자음 18개, 중성 모음 21개, 종성 자음 7개 및 운율구에서 확실한 경계가 있는 구간에 해당하는 마침표와 약한 경계의 쉼표, 그리고 어절구간의 경계를 나누는 blank도 하나의 음소로 하여 총16비트의 신경망의 입력패턴으로 사용하여 운율을 발생시키고자 한다.

Ⅲ-1 입력패턴

먼저 분절요인에 의한 영향을 고려해주기 위해 전후 3음소를 동시에 입력시키고, 전처리로써 운율구에 대한 각 단어를 초, 중, 종성 및 경계구간을 분리하여 문-음소변환 알고리듬을 사용하여 음운 변동을 적용한 후 이를 통해 얻어진 음소, 즉 자음 18개, 중성 모음 21개, 종성 자음 7개 및 경계구간을 나타내는 마침표와 쉼표,

그리고 blank(□)도 하나의 음소로 하여 6 bit로 할당하고, 운율구내에서의 음소의 상대적인 위치에 관한 정보와 종음소에 대한 정보를 각각 5 bit로 할당하여, 총16비트의 신경망의 입력 패턴으로 구성하였다.

다음은 아래 문장에서 예로 /□장난/에 대한 각 음소 /□/ /ㅈ/, /ㅏ/, /ㅇ/, /ㄴ/, /ㅏ/ /ㄴ/에 대해 1진 벡터로 표현한 예이다.

"철수는 장난감을 사고, 영희는 과자를 샀다.

문장내의 /□장난/에 대한 입력패턴 예

□[110000 01001 11011]
ㅈ[001001 01010 11011]
ㅏ[010100 01011 11011]
ㅇ[101111 01100 11011]
ㄴ[101010 01101 11011]
ㅏ[010100 01110 11011]
ㄴ[101010 01111 11011]

이렇게 음소 기호열로 구성된 입력패턴은 총 112개의 비트열로 구성되는데, 입력층에 매 패턴제시마다 전후 3음소씩 모두 7음소가 학습에 참여하게 된다. 학습에 참여하는 7개의 음소 중 실제로 학습이 이루어지는 음소는 네 번째 음소이며, 각각의 음소들은 순차적으로 쉬프트 되면서 학습이 이루어지게 된다.

Ⅲ-2. 출력패턴

각 음소의 지속 시간 동안에 피치와 에너지 변화에 대한 출력 패턴은 회귀 분석을 통한 추세선을 이용하여 3차 다항식으로 근사화한 후 그 계수인 p_3, p_2, $p_1(e_3, e_2, e_1)$과 초기 피치값 $p_0(e_0)$를 1진 벡터화 하여 작성하였다. 그리고 다항식의 변수인 지속 시간 d은 그 프레임수를 2진부호화 하여 출력 패턴을 작성하고, 무성자음의 경우 피치가 존재하지 않으므로 피치변화곡선의 계수와 초기 피치값은 모두 0으로 부호화하였으며 지속시간은 모음과 유성자음의 경우와 마찬가지로 그 프레임수를 1진 벡터화하여 작성한다.

이때 각 음소의 지속시간과 초기 피치 및 에너지값의 범위는 각각 0~300 msec (0~24 frame), 0~6.0로 하고 이를 1진 벡터화하기 위해서 각각 41bit씩을 할당하였다. 운율 변화 곡선의 다항식 계수들은 그 최대 존재 범위가 0~9이므로 이를 부호화하기 위해서 지속시간, 초기 피치 및 에너지값과 마찬가지로 41 bit를 할당하여 첫 1 bit는 부호, 다음 10 bit는 단 자리를 다음 10 bit는 소숫점 첫째 자리를 다음 10 bit 는 소숫점 둘째 자리를 그리고, 다음 10 bit는 소숫점 셋째 자리를 나타내도록 한다[4].

Ⅳ. 결론

자연성이 부가된 문장단위의 합성음에 대한 음성데이타의 효율적 분석을 위해서는 각 문장의 내부를 음성적으로 의미있는 단위로 끊어주는 것이다. 여러개의 낱말이 모여 구를 이루고 이들이 다시 절로 그리고 문장으로 묶어져 나가는 것과 같이 문장단위상에 음성데이타에서도 끊어읽기 및 breath group의 과정을 통하여 나름대로의 단위를 만들어 나가는 단위구성의 방법이 존재한다. 운율구의 단위는 화자의 입장에서 끊어읽기, 끊김의 청각적 지각을 이용하여 만들어지며 호름이 끊이지는 부분이 운율구의 경계가 된다.

본 논문은 단어의 음소성분과 문장의 마침표와 같은 확실한 경계가 있는 구간과 쉼표와 같은 약한 경계, 그리고 blank와 같은 어절구간의 경계를 나누어서, 이것도 단어의 음소의 변화성분과 마찬가지로 음소로 취급하여 신경망을 학습시키고, 해당음소의 운율구내에서의 상대위치에 따른 운율 변화도 같이 훈련시켜, 복잡한 운율법칙 알고리듬을 위한 프로그램 작성없이 연속음의 운율변화곡선을 발생시키는데 필요한 신경망의 입력패턴의 방법을 제안하였다.

[참 고 문 헌]

[1] Eric Sanders and Paul Taylor, "Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis." in EU ROSPEECH'95 Spain, 1995.

[2] 임 운천, 한국어 법칙합성을 위한 운율법칙 구현에 관한 연구, 서울대학교 박사학위논문, 1991.

[3] 성철재, "한국어 리듬의 실험음성학적 연구", 서울대학교 박사논문, 1995.

[4] 류창수, "신경망 합성에 따른 운율 제어기 성능 비교에 관한 연구", 호서대학교 석사논문, 1998

[5] 김현준, "신경망을 사용한 문-변이음 변환에 관한 연구", 호서대학교 석사논문, 1998.

[6] 김선철, "국어 억양의 음성학·음운론적 연구", 서울대학교 박사논문 1996.

[7] 김연준, 오영환 "한국어 문서-음성 변환 시스템에서의 구문분석에 의한 운율조절에 관한 연구" 제 10회 음성통신 및 신호처리 워크샵 논문집, 1993.

[8] 허 준, 무제한 단어 한국어 음성합성 시스템에서의 운율정보 구현에 관한 연구, 서울대학교 석사학위 논문, 1990.

[9] Ostendorf, "Parse scoring with prosodic information : an analysis and synthesis approach." in

Computer Speech and Language. July 1993.
[10] 이현복, "음성학과 언어학", 서울대학교출판부,1996
[11] 정국외 4, "음성인식/합성을 위한 국어의 음성-음운론적 특성 연구" 한국 음향학회지 제 13권 6호,1994.
[12] J. Allen, M. S. Hunnicutt & D. Klatt , From Text To Speech : The MITalk System.Cambridge University Press, 1987.
[13] D. H. Klatt, "Structure of phonological rule component for synthesis by rule program", IEEE Vol.ASSP-24 No.5, pp.391-398, 1976.
[14] D. O'Shaughnessy, "Automatic Speech Synthesis", IEEE Communication magazine, pp26-34, 1983. 12.
[15] Do-Heung Ko, Declarative intonation in Korean ; An acoustical study of F0 declination, Ph. D dissertation, Univ. of Kansas, 1988.
[16] N. Umeda, "Linguistic rules for text-to-speech synthesis", Proc. of IEEE, vol. 64, No. 4, pp. 433-451, Apr. 1976.
[17] R. P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magagine, Vol. 4, No. 2, pp. 4-22, April 1987.
[18] J. M. Zurada, Introduction to Artificial Neural Systems, West Publishing Company, 1992.