1999년도 한국음향학회 학술발표대회 논문집 제18권 제1(s)호

신경망을 이용한 한국어 운율 발생에 관한 연구

A Study on the Prosody Generation of Korean Sentences using Neural Networks

이일구* 민경중* 강찬구** 임운천*
*호서대학교 전자공학과, **안양과학대

Il-Goo Lee*, Kyoung-Joong Min, Chan-Koo Kang, Un-Cheon Lim
*Dept. of Electronics Eng., Hoseo University
**Anyang Science University

E-mail: uclim@dogsuri.hoseo.ac.kr kjmin@dogsuri.hoseo.ac.kr

요약

합성단위, 합성기, 합성방식 등에 따라 여러 가지 다양한 음성합성시스템이 있으나 순수한 법칙합성 시스템이 아니고 기본 합성단위를 연결하여 합성음을 발생시키는 연결합성 시스템은 연결단위사이의 매끄러운 합성계수의 변화를 구현하지 못해 자연감이 떨어지는 실정이다. 자연음에 존재하는 운율법칙을 정확히 구현하면 합성음의 자연감을 높일 수 있으나 존재하는 모든 운율법칙을 추출하기 위해서는 방대한 분량의 언어자료 구축이 필요하다.

일반 의미 문장으로부터 운율법칙을 추출하는 것이 바람직하겠으나, 모든 운율 현상이 포함된 언어자료는 그 문장 수가 극히 방대하여 처리하기 힘들기 때문에 가능하면 문장 수를 줄이면서 다양한 운율 현상을 포함하는 문장 군을 구축하는 것이 중요하다.

본 논문에서는 음성학적으로 균형 잡힌 고립단어 412단어를 기반으로 의미문장들을 만들었다. 이들 단어를 각 그룹으로 구분하여 각 그룹에서 추출한 단어들을 조합시켜 의미 문장을 만들도록 하였다. 의미 문장을 만들기 위해 단어 목록에 없는 단어를 첨가하였다. 단어의 문장 내에서의 상대위치에 따른 운율 변화를 살펴보기 위해 각 문장의 변형을 만들어 언어자료에 포함시켰다.

자연감을 높이기 위해 구축된 언어자료를 바탕으로 음성데이타베이스를 작성하여 운율분석을 통해 신경망을 훈련시키기 위한 목표패턴을 작성하였다. 문장의 음소열을 입력으로 하고 특정음소의 운율정보를 발생시키는 신경망을 구성하여 언어자료를 기반으로 작성한 목표패턴을 이용해 신경망을 훈련시켰다. 신경망의 입력패턴은 문장의 음소열 중 11개 음소열로 구성된다. 이 중 가운데 음소의 운율정보가 출력으로 나타난다. 분절요인에 의한 영향을 고려해주기 위해 전후 5음소를 동시에 입력시키고 문장내에서의 구문론적인 영향을 고려해주기 위해 해당 음소의 문장내에서의 위치, 운율구에 관한 정보등을 신경망의 입력 패턴으로 구성하였다. 특정화자로 하여금 언어자료를 발성하게 한 음성자료의 운율정보를 추출하여 신경망을 훈련시킨 결과 자연음의 운율과 유사한 합성음의 운율을 발생시켰다.

I. 서 론

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

법칙 합성 시스템은 합성단위, 합성기, 합성방식등에 따라 여러 가지 다양한 시스템이 있으나 순수한 법칙합성 시스템이 아니고 기본 합성단위를 연결하여 합성음을 발생시키는 연결합성 시스템은 연결단위사이에서의

매끄러운 합성계수의 변화를 구현하지 못해 자연감이 떨어지는 실정이다. 특히 시간영역 법칙합성 시스템의 합성음은 이해도는 향상되었음에도 불구하고 자연감이 많이 떨어지고 있다[8].

음성 언어자료는 상태에 따라 많은 데이터량을 가지고 있어, 이러한 데이터베이스를 신경망에 학습시키는 데에는 많은 처리과정 및 시간이 필요하게 된다. 본 논문에서는 언어자료를 음성학적으로 균형 잡힌 고립단어 412 단어를 기반으로 의미문장들을 만들었다. 이들 단어를 각 그룹으로 구분하여 각 그룹에서 추출한 단어들을 조합시켜 의미 문장을 만들도록 하였다. 명사형 단어(262 단어)와 동사형 단어(106 단어) 그리고 형용사, 부사, 감탄사 등의 단어(44 단어)로 구성된 단어 목록을 이용하여 의미문장을 작성하였고, 의미 문장을 만들기 위해 단어 목록에 없는 단어를 첨가하였다. 단어의 문장 내에서의 상대위치에 따른 운율 변화를 살펴보기 위해 각 문장의 변형을 만들어 언어자료에 포함시켰다. 이를 토대로 음성시료를 채집하였고 운율분석을 하여 신경망 훈련에 필요한 목표패턴을 구축하여 신경망을 학습시킴으로써 연속음에서 좀더 이해도 및 자연성을 향상시키는 방법을 제안하고자 한다.

Ⅱ. 운율제어

일반적으로 서로간 자연스러운 대화라든가 글을 읽을 때의 음성, 즉 자연음은 화자의 감정상태, 말의 내용 또는 강세, 발음속도와 같은 의미론적인 정보와 전체의 구문구조와 문장내에서의 구와 절의 경계위치, 기능, 상호 결합관계 등의 구문론적인 정보, 단어에서의 강세유형과 각 분절음소의 전후 결합에 의한 영향이 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.

피치의 변화에 영향을 주는 요인으로 먼저 의미론적인 요소 즉 대비, 강조, 화자의 감정상태와 발음속도 등이 있고, 구문론적인 측면에서는 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.

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]. 한국어의 경우 전체 문장 길이의 분산이 각 음절길이의 분산의 합에 비해 적어 음절이 기본단위가 아니고 분절이 기본단위임을 제시하고 있다.

에너지는 음소를 구별하는데 중요한 운율요소 중 하나로, 합성음의 자연성에 미치는 영향이 큰 강세 및 억양에 의해 변화하며 운율법칙의 중요한 요소이다[5].

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.

Ⅲ. 언어자료 구축 및 음성시료 채집

Ⅲ-1 언어 자료 구축

음성 언어 처리를 위해 분석용, 합성용, 인식용, 평가용 음성자료를 구축하여 음성 데이터베이스를 구축하는 것이 기본이다. 운율 법칙을 훈련시키기 위한 언어자료는 일반 문장으로 구성된 언어자료, 전달 문구로 구성된 언어자료, 무의미 문장으로 구성된 언어자료, reiterant speech로 구성된 언어자료 등을 고려하여 사전에 운율이 어떻게 변하는지 살펴보는 것이다. 존재하는 모든 운율 법칙을 추출하기 위해서 방대한 분량의 언어자료가 구축된다. 실제 한국어의 경우 시스템공학연구소의 국어공학센터에서 100만 어절의 한국어 텍스트 코퍼스를 구축하였다. 여기에는 소설, 산문, 저술 등에서 추출한 문장들이 포함되어 있다. 이 방대한 언어자료를 녹음하여 분석하는 것은 거의 불가능하기 때문에 이 중에 일부를 균형 있게 채택하여 문장 수를 줄여 실험에 이용하고 있다. 실제 코퍼스 중 1만 문장을 대상으로 음성학적 균형이 잡힌 589 문장을 엄선하여 음성 데이터 베이스를 구축하는 사례도 보고되었다. 이 같이 일반 문장을 분석자료로 사용하기 위해서는 엄청난 양의 자료 분석을 수반하기 때문에 단기간에 처리할 수 없어 대개 제한된 개수의 문장을 실험에 이용하게 된다. 문장이 제한되면 모든 운율법칙이 포함된다고 보장할 수 없다는 단점이 있다.

본 논문에서는 음성학적으로 균형 잡힌 고립단어 412 단어를 기반으로 의미문장들을 만들었다. 이들 단어를 각 그룹으로 구분하여 각 그룹에서 추출한 단어들을 조합시켜 의미 문장을 만들도록 하였다. 명사형 단어(262 단어)와 동사형 단어(106 단어) 그리고 형용사, 부

사, 감탄사 등의 단어(44 단어)로 구성된 단어 목록을 이용하여 의미문장을 작성하였다. 의미 문장을 만들기 위해 단어 목록에 없는 단어를 첨가하였다. 단어의 문장 내에서의 상대위치에 따른 운율 변화를 살펴보기 위해 각 문장의 변형을 만들어 언어자료에 포함시켰다.

Ⅲ-2 음성 시료 채집

운율자료를 추출하기 위해 언어자료를 구축하였고 이것을 특정화자로 하여금 발성하게 하였다. 화자는 표준말을 쓰고 발음을 정확히 하는 남성화자로 선정하였다.

DAT용 마이크를 이용하여 언어자료를 화자로 하여금 발성하게 하여 DAT로 녹음하였다. 표준 주파수는 추후 과정을 거쳐 11.025Khz로 조절하고 표본당 비트수는 16비트로 지정하였다.

언어자료를 토대로 화자로 하여금 무반향실에서 발음하도록 하고 이것을 녹음하였고 DATLink를 이용해 컴퓨터로 음성시료를 옮겨 편집을 통해 불필요한 부분을 제거하여 음성데이터 규모를 최소로 줄이고 편집과정에서 발견된 녹음 불량 문장은 다시 발음하게 하여 음성데이터 베이스를 구축하였다. 그림 1은 문장단위 음성파형이다.

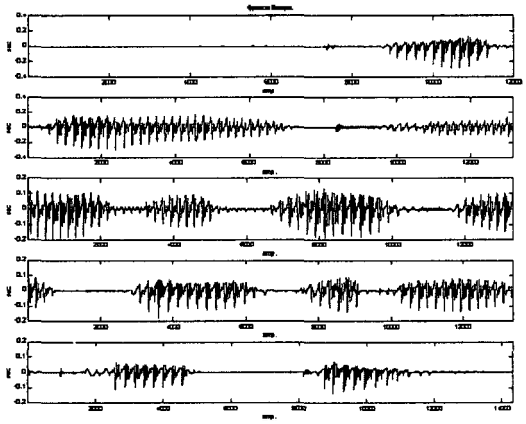

그림 1. 음성 표본
Fig. 1. Speech Sample

Ⅳ. 신경망 구성 및 훈련

신경망 시스템은 계층적 구조인 입력층, 은닉층, 출력층의 3개의 층으로 구성하였다. 입력층은 11개의 음소열를 나타내는 유니트들로 이루어져 있는데 한 음소당 8bit의 유닛으로 지정하여 입력층은 총 88개의 입력 유닛으로 구성하였다. 각종 규칙을 적용한 언어자료를 가지고, 초성 자음 18개, 중성 모음 21개, 종성 자음 8개 및 마침표, 쉼표, 그리고 blank를 음소로 하여 신경망을 훈련시켜 자연음의 운율과 유사한 합성운율을 발생시킨다.

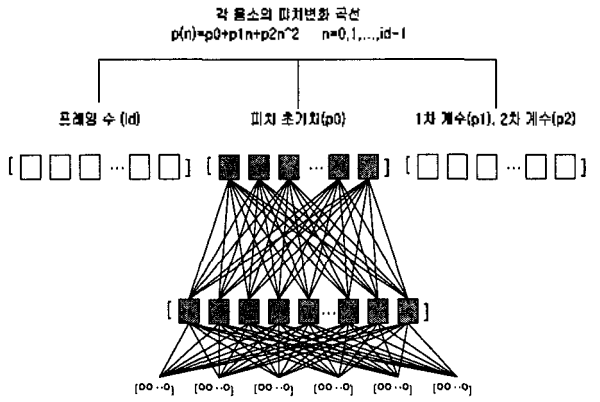

그림 2. 피치 신경망의 블록도
Fig. 2. Block diagram of pitch neural networks

Ⅳ-1 입력패턴

먼저 분절요인에 의한 영향을 고려해주기 위해 전후 5음소를 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 초성자음 18개와 중성모음 21개, 종성자음 8개, 구두점 마침표와 쉼표 각 1개, 띄어쓰기 부호를 일반 공백에 1개, 운율 경계로 사용되는 공백에 1개를 지정하고, 문장 시작 부분을 알리는 부호 1개를 포함하면 총 51개의 심볼이 필요하다. 51가지 정보를 2진수로 표시하기 위해 6bit를 할당하면 된다. 한 음소당 8bit를 할당하여 2bit는 추후 4가지 정보를 추가로 사용할 수 있게 하였다.

Ⅳ-2 출력패턴

각 음소의 지속 시간 동안에 피치와 에너지 변화에 대한 출력 패턴은 회귀 분석을 통한 추세선을 이용하여 2차 다항식으로 각 변화곡선을 근사하였다. 먼저 피치 변화 곡선은 분절의 전체 프레임 수와 피치 주기로 결정되는데 이를 다항식으로 근사하면

$$p(n) = p_0 + p_1 n + p_2 n^2 \quad n = 0, \cdots, id-1 \qquad 식(1)$$

id는 각 분절의 지속시간(프레임 수)이고 p_0는 초기 피치 주기를 나타내며 p_1과 p_2가 다항식의 계수가 된다. 그림 3은 모음 '예'의 피치 변화곡선과 추세선을 나타낸 것이다.

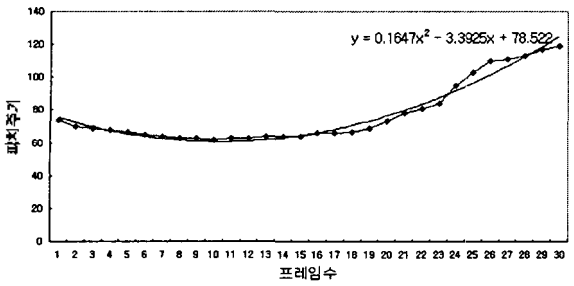

그림 3. 피치 변화곡선과 그추세선
Fig. 3. Pitch contour and it's regressive line

에너지 변화 곡선도 마찬가지로 에너지 초기치와 2차 다항식 계수로 다음과 같이 근사할 수 있다.

$$e(n) = e_0 + e_1 n + e_2 n^2 \quad n = 0, 1 \cdots, id-1 \qquad 식(2)$$

여기서 e_0는 초기 에너지이고 id는 분절의 지속시간이다. e_1, e_2는 2차 다항식 계수이다. 그림 4는 모음 '에'의 에너지 변화곡선과 추세선을 나타낸 것이다.

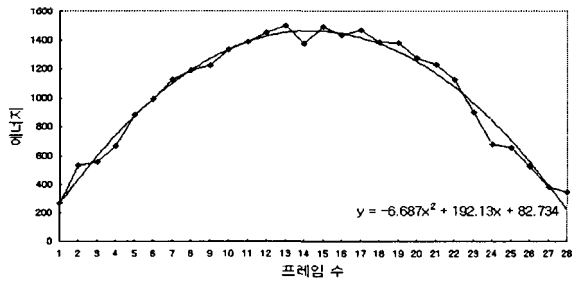

그림 4. 에너지 변화곡선과 그 추세선
Fig. 4. An Energy contour and it's regressive line

이 계수인 p_2, $p_1(e_2, e_1)$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 백터화하여 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].

초기 피치값은 모두 0으로 부호화하였으며 지속시간은 모음과 유성자음의 경우와 마찬가지로 그 프레임수를 1진 백터화하여 작성한다. 이때 각 음소의 지속시간과 초기 피치 및 에너지값의 범위는 각각 0~300 msec (0~24 frame), 0~6.0로 하고 이를 1진 백터화하기 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].

V. 결론

자연성이 부가된 문장단위의 합성음에 대한 음성데이타의 효율적 분석을 위해서는 각 문장의 내부를 음성적으로 의미있는 단위로 끊어주고, 개별적 언어자료구축에 따른 중복투자를 줄이고, 각종 알고리즘을 적절히 비교 평가하기 위해서 데이터량을 줄이면서 실제 한국어의 발성에 나타나는 음운현상을 가능한 많이 포함하며, 특정 태스크에 집중되지 않는 것이 바람직하다. 음성 언어자료는 상태에 따라 많은 데이터량을 가지고 있어, 이러한 데이터베이스를 신경망에 학습시키는 데에는 많은 처리과정 및 시간이 필요하게 된다.

본 논문에서는 언어자료를 음성학적으로 균형 잡힌 고립단어 412 단어를 기반으로 의미문장들을 만들었다. 이들 단어를 각 그룹으로 구분하여 각 그룹에서 추출한 단어들을 조합시켜 의미 문장을 만들도록 하였다. 명사형 단어(262 단어)와 동사형 단어(106 단어) 그리고 형용사, 부사, 감탄사 등의 단어(44 단어)로 구성된 단어 목록을 이용하여 의미문장을 작성하였다 의미 문장을 만들기 위해 단어 목록에 없는 단어를 첨가하였다. 단어의 문장 내에서의 상대위치에 따른 운율 변화를 살펴보기 위해 각 문장의 변형을 만들어 언어자료에 포함시켰다. 이를 토대로 음성시료를 채집하였다. 신경망을 이용해 자연음으로부터 추출한 운율 정보를 학습시켜, 문장의 음소열이 입력으로 들어오면 해당하는 음소의 운율 정보를 출력하도록 하여, 자연스러운 운율 변화를 보이는 운율 발생기를 구현하였다. 신경망의 학습단계에서의 추정율은 90% 이상이었고, 평가 단계에서의 추정율도 89%이상이 되었다.

[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] 민경중, 이일구, 강찬구, 임운천, "한국어 문장단위 운율 발생에 관한 연구," 1998년도 한국 음향학회 학술발표대회 논문집 제 17권 2(s)호, pp. 419-423.

[10] 민경중, 임운천, "문장 단위 운율제어를 위한 신경망의 입력 패턴에 관한 연구," 제 15회 음성 통신 및 신호처리 워크샵 논문집, KSCSP'98 Vol.15 NO. 1, pp.105-108.

[11] Ostendorf, "Parse Scoring with Prosodic Information : an Analysis and Synthesis Approach." in Computer Speech and Language. July 1993.

[12] 이현복, "음성학과 언어학", 서울대학교출판부,1996

[13] 정국외 4, "음성인식/합성을 위한 국어의 음성-음운론적 특성 연구" 한국 음향학회지 제 13권 6호,1994.

[14] J. Allen, M. S. Hunnicutt & D. Klatt , From Text To Speech : *The MITalk System. Cambridge University Press*, 1987.

[15] D. H. Klatt, "Structure of Phonological Rule Component for Synthesis by Rule Program", IEEE Vol.ASSP-24 No.5, pp.391-398, 1976.

[16] D. O'Shaughnessy, "Automatic Speech Synthesis", IEEE Communication magazine, pp. 26-34, 1983.

[17] Do-Heung Ko, Declarative Intonation in Korean ; An Acoustical Study of F0 Declination, Ph. D dissertation, Univ. of Kansas, 1988.

[18] N. Umeda, "Linguistic Rules for Text-to-speech Synthesis", Proc. of IEEE, vol. 64, No. 4, pp. 433-451, Apr. 1976.

[19] R. P. Lippmann, "An Introduction to Computing with Neural Nets", IEEE ASSP Magagine, Vol. 4, No. 2, pp. 4-22, April 1987.

[20] J. M. Zurada, *Introduction to Artificial Neural Systems*, West Publishing Company, 1992.

[21] 허웅, 국어 운운학, 정음사, 1985

[22] H. Dettwealer and W. Hess, "Concatenation Rules for Demisyllable Speech Synthesis," NATO ASI Series, vol. F16, 1985