1999년도 한국음향학회 학술발표대회 논문집 제18권 제2(s)호

인공 신경망을 이용한 한국어 문장단위 운율 발생에 관한 연구

A study on the Prosody Generation of Korean Sentences using Artificial Neural networks

이일구, 민경중, 강찬구*, 임운천
Dept. of Electronics Eng., Graduate School, Hoseo Univ.
Anyang Technical College.*
uclim@mail.hoseo.ac.kr

요약

TTS(Text-To-Speech) 시스템 합성음성의 자연감을 개선하기 위해 하나의 언어에 대해 존재하는 운율 법칙을 정확히 구현해야 한다. 존재하는 운율 법칙을 추출하기 위해서는 방대한 분량의 언어 자료 구축이 필요하다. 그러나 이 방법은 존재하는 운율 현상이 포함된 언어자료에 대해 완벽한 운율을 파악할 수 없으므로 합성음성의 질을 좋게 할 수 없다.

본 논문은 한국어 음성의 운율을 학습하기 위해 2개의 인공 신경망을 제안한다. 하나의 신경망으로 문장의 각 음소에 대한 피치 변화를 학습시키는 것이며, 다른 하나는 에너지 변화를 학습하도록 하였다.

신경망은 BP 신경망을 이용하며 11개의 음소를 나타내기 위해 11개의 입력과, 중간 음소의 피치와 에너지 변화곡선을 근사하는 다항식 계수를 출력하도록 하였다.

신경망시스템의 학습과 평가에 앞서, 음성학적 균형잡힌 고립단어를 기반으로 의미있는 문장을 구성하였다. 문장을 남자 화자로 하여금 읽게 하고 녹음하여 음성 DB를 구축하였다. 음성 DB에 대해 각 음소의 운율 정보를 수집하여 신경망에 맞는 목표 패턴과 훈련 패턴을 작성하였다. 이 목표 패턴은 회귀분석을 통한 추세선을 이용해 피치와 에너지에 대한 2차 다항식계수로 구성하였다.

본 논문은 목표패턴에 맞는 신경망을 학습시켜 좋은 결과를 얻었다.

I. 서 론

음성은 인간과 기계간에 정보전달에 가장 바람직한 통신의 도구중 하나이다. 기계가 인간 음성을 이해하고, 합성된 음성을 발생시키는 것이다. 음성 합성 기술은 장난감에서 TTS 시스템까지 응용되어진다. 특히, TTS 시스템의 음성합성의 목표는 무제한 단어와 자연감을 향상시키는 것이다.

1970년 중반, TTS 합성은 디지털 신호 처리에 대한 디지털 계산 능력이 향상됨에 따라 많은 연구가 되어졌다. 좀더 향상된 이해도와 자연성을 가진 TTS 시스템으로 만들려는 다각적인 방법이 논의되었다. TTS 시스템의 이해력과 자연감을 향상시키기 위해 언어에 대해 정확한 음향-음성학적 정보 뿐만 아니라 운율 정보가 필요하다. TTS 시스템에 대한 합성된 음성의 질은 인간의 음성보다 떨어졌다. 그러므로 시스템은 음성 질을 높이기 위해 운율 법칙을 적용하였으며 그 방법은 자연음로부터 얻거나 각 언어에 대해 음성학적 정보를 이용 하였다. 그러나 충족된 운율 법칙은 자연음의 모든 운율 법칙을 반영하는데 불충분하고, 법칙화하는데 어려움이 따른다. 그래서 운율 법칙에 대한 불정확한 표현은 합성의 음성의 질을 낮추어지는 결과를 가져왔다.

본 논문은 신경망으로 하여금 자연음으로 부터 운율 법칙을 학습시키며, 문장내의 위치와 음소에 대한 입력열로 하고, 출력은 문장내의 음소에 대한 운율정보로 한다.

신경망의 훈련시키기 위해 한국어 문장으로 구성된 자료로 구축하고, 문장의 각 단어는 음성학적으로 균형잡힌 고립단어 자료로 구성하였다. 단어로 구성된 문장을 남자 화자 1인으로 하여금 읽게 하여 녹음하였고, 음성 DB를 구축하였다. 문장내의 각 음소의 운율 모양을 분석하여 각 패턴을 만들었다. 그것을 이용하여 학습단계의 학습패턴으로 사용하고, 또한 훈련단계의 훈련패턴으로 사용하였다. 신경망의 입력 패턴은 11개의 음소로 구성하였다.

II. 한국어에서의 운율

운율에는 세 가지 요소로서 지속시간, 크기 그리고 피치로 주어진다. 각 음소의 이들 파라미터는 의미론적 정보와 구문론적 정보로 나눠진다.

의미론적 정보는 화자의 감정, 강세, 말의 내용 그리고 발음속도로 나눠진다. 하지만 이 모든 변화에 대한 법칙을 만들어내기 어렵다. 본 논문은 한국어 문장 내에서의 운율 훈련을 쉽게하기 위해 특정한 상태로 제한했다. 남자 화자는 어떤 감정이 없이 평상속도로 문장을 읽게 하였다.

본 논문은 운율을 구문론적 정보, 의미론적 정보 그리고 음절의 정보로 표시했다. 의미적 강세는 단어의 패턴의 중요한 법칙이며 문장에서의 강세는 좀더 중요하다. 각각에 대한 부분의 피치, 지속시간, 크기에의 변화는 연속된 부분의 상호작용의 결과이므로 이 문제가 좀더 연구되어야 한다.

III. 신경망

신경망은 자연음에서 운율법칙을 학습시켜 학습 운율 법칙에 대한 알고리즘을 만들 필요가 없고, 부정확한 운율 법칙에 의한 합성음성 음질을 걱정할 필요가 없다. 신경망은 연산 접근에 의해 정의되지 않고 학습 할 수 있으며, 충분한 음성 데이터를 가지고 있는 신경망을 학습시키면 연산 시스템의 수행결과보다 자연적인 합성 음성을 얻게 된다.

본 논문은 BP(Back Propagation) 신경망을 문장내의 각각 음성의 운율 법칙으로 학습시켰다. 그림 1은 하나의 은닉층을 가진 전형적인 BP 신경망이다. 이 신경망은 입력 음소 열의 중간 음소의 피치 변화를 훈련시키는 그림이다.

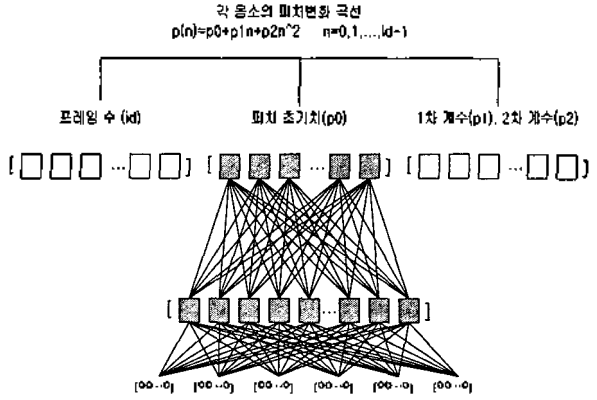

그림 1. 피치 신경망의 블록도

Fig. 1. Block diagram of pitch neural networks

각종 규칙을 적용한 언어자료를 가지고, 초성 자음 18개, 중성 모음 21개, 종성 자음 8개 및 마침표, 쉼표, 그리고 blank를 음소로 하였다. 이 정보를 2진수로 표시하면 6bit로 되므로, 한 음소당 8bit를 할당하였다.

운율구내의 음소의 수를 2에서 10개로 하고, 전후 음소 그리고 마지막 운율의 길이의 정보를 위해 신경망의 입력수는 11음소로 정의했다. 음소중에 6번째 음소 즉 중간 음소의 운율 정보를 신경망의 출력값으로 사용한다. 운율 정보를 학습하기 위해 하나의 은닉층의 구조로 선택하였고, 그리고 은닉층의 수는 입력수의 수와 같게 하였다.

신경망의 출력은 입력 음소열열 중 중간에 해당하는 음소의 운율 정보 즉 피치와 에너지 변화에 대한 근사다항식 계수와 초기치로 구성된다.

음성 데이터를 프레임별로 분석한 실제 음성의 스펙트럼과 운율 변화를 나타내고, 샘플링 주파수는 10 KHz, 한 프레임을 256 샘플, 그리고 중첩은 128 샘플로 하고, 프레임간의 시간 간격은 12.8 msec로 하였다. 각 음소의 프레임의 수는 1에서

24 frames로 정했다. 각 음소의 피치와 에너지는 회귀분석을 통한 추세선을 이용하여 추출하였다.

다음 식은 각 음소의 피치와 에너지의 2차 다항식 근사하였다.

$$p(n) = p_2 \times n^2 + p_1 \times n + p_0 \quad (1)$$

$$e(n) = e_2 \times n^2 + e_1 \times n + e_0 \quad (2)$$

$$0 \le n \le d-1$$

d는 각 분절의 지속시간(프레임 수)이고, p2과 p1는 다항식의 계수이며 p0은 피치 초기값을 나타낸다. 그리고 e2, e1은 에너지의 다항식의 계수이며, e0는 에너지 초기값을 나타낸다.

그림 2는 모음 '에'의 피치 변화곡선과 추세선을 나타내며, 그림 3는 모음 '에'의 에너지 변화곡선과 추세선을 나타낸다.

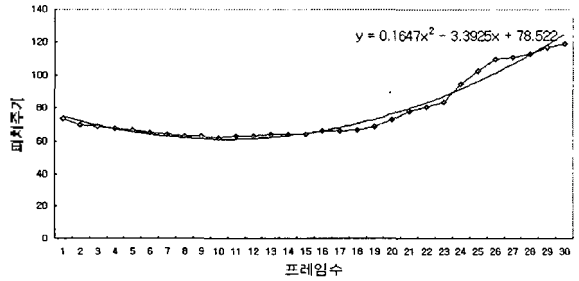

그림 2. 피치 변화곡선과 그 추세선

Fig. 2. A pitch contour and it's regressive line

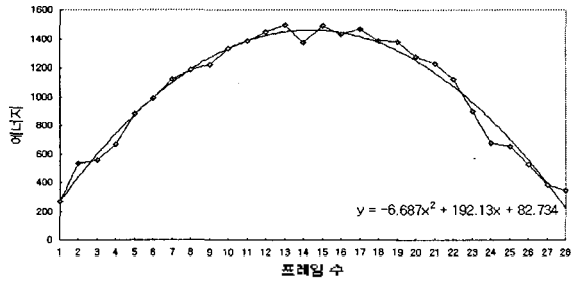

그림 3. 에너지 변화곡선과 그 추세선

Fig. 3. An Energy contour and it's regressive line

각 음소의 출력과 평가는 남자화자 1인 구현한 음성 DB를 단기 분석 알고리즘을 이용하여 분석하였고, 각 음소의 피치와 에너지의 다항식 계수와 초기 값을 부호화 하여 구했다.

IV. 실험

본 논문은 음성학적으로 균형잡힌 고립단어 412단어를 기반으로 의미문장들을 만들었다. 이들 단어를 각 그룹으로 구분하여 각 그룹에서 추출한 단어들을 조합시켜 의미 문장을 만들도록 하였다.

화자는 남성화자로 선정하고 문장을 세 번씩 읽도록 하였다. 음성 DB는 단기 분석 알고리즘을 이용하여 분석하였다. 그리고 목표 패턴은 운율 정보를 분석하여 2차 다항식으로 운율곡선을 근사하여 구한 다항식 계수와 초기치로 작성하였다.

2회 발성한 음성자료에서 구한 목표패턴에 대해 신경망을 학습시켰다. 임계치를 정해 200회 이전에 학습오차가 임계치 이하로 떨어지면 다음 패턴을 학습하도록 하였다.

표 1은 학습 단계에서의 신경망의 추정율이다.

표 1. 학습단계의 추정율

d	p0	p1	p2
92.4%	91.7%	92.5%	92.2%

d	e0	e1	e2
90.4%	91.7%	91.3%	90.3%

훈련 단계에서는 신경망의 출력패턴을 학습 패턴과 비교하여 세 번째 자연음에서 추출하여 학습단계의 추정률과 비슷함을 비교하였다. 표2는 학습단계의 신경망의 추정율이다.

표 2. 훈련 단계의 추정율

d	p0	p1	p2
90.3%	90.3%	89.4%	90.7%

d	e0	e1	e2
88.4%	90.1%	89.4%	88.3%

V. 결 과

피치에 대한 신경망 학습 단계에서의 추정율은 92%, 평가 단계에서는 90%로 나타나고, 에너지에 대한 신경망의 학습 단계에서의 추정율은 90%, 평가 단계에서는 89% 나타났다.

신경망은 자연음로부터 운율정보를 얻어 훈련시키고, 음소수는 분절의 영향과 단어의 액센트 영향을 고려하여 11음소로 하였다. 한국어 운율구의 음소수가 11개가 넘는 경우가 있으므로 이런 경우 운율구내에서의 음소의 위치에 따른 운율 변화를 신경망에 학습할 수 없으므로 입력음소 수에 대한 연구가 더 필요하다.

[참 고 문 헌]

[1] Eric Sanders and Paul Taylor, "Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis." in EUROSPEECH'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] 김연준, 오영환, "한국어 문서-음성 변환 시스템에서의 구문분석에 의한 운율조절에 관한 연구", 제 10회 음성통신 및 신호처리 워크샵 논문집, 1993.

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

[8] 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.

[9] 이일구, 민경중, 강찬구, 임운천, "신경망을 이용한 한국어 운율 발생에 관한 연구", 1999년도 한국 음향학회학술대회 논문집 제 18권 1(s)호, pp. 65-69

[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, pp. 26-34, 1983.

[15] Sok-Wang Chang, Hyung-Joon Kim, Chang-Su Ryoo, Un-Cheon Lim, "A Study on the Prosody Generation in Isolated Words with an Artificial Neural Network", ICSP 97', pp207-211, 1997

[16] Hyun-Joon Kim, Sok-Wang Chang, Chang-Su Ryoo, Un-Cheon Lim, " A Study on the Prosodic Marker in a Korean Sentence", ICSP 97', pp213-218, 1997

[17] Kyung-Joong Min, Joon-Sik Kim, Un-Cheon Lim, "Input/Output Pattern of Neural Networks for Prosody Generation of Korean Sentence", ICSP 99', pp161-165, 1999

[18] Il-Goo Lee, Chan-Goo Kang, Joon-Sik Kim, Un-Cheon Lim, "Prosody Generator for Speech Synthesizer Using Artificial Neural Networks", ICSP 99', pp183-186, 1999

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

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

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

[22] 허웅, 국어 운운학, 정음사, 1985