2000년도 한국음향학회 학술발표대회 논문집 제19권 제1(s)호

인공신경망의 운률 발생에 관한 연구

A Study on the prosody generation of artificial neural networks

신동엽, 민경중, 강찬구, 임운천
호서대학교 대학원 전자공학과
336-795. 충청남도 아산시 배방면 세출리 산 29-1
uclim@office.hoseo.ac.kr

요약

문-음성 합성기의 자연감을 높이기 위해 주로 자연음에 존재하는 운률 법칙을 정확히 구현해 주어야 한다. 일반적으로 언어학적 정보를 이용하거나 자연음으로부터 추출한 운률 정보를 추출한 운률 법칙을 합성에 이용하고 있다.

이와 같이 구한 운률 법칙이 자연음에 존재하는 모든 운률 법칙을 포함할 수 있으면, 자연스러운 합성음을 들을 수 있겠으나, 실질적으로는 모든 법칙을 구현한다는 것은 어려운 실정이고, 자연음으로부터 추출한 운률 법칙이 잘못 구현되는 경우 합성음의 자연성이 떨어지는 것을 피할 수 없을 것이다.

이런 점을 고려하여 우리는 자연음에 내재하는 운율 법칙을 훈련을 통해 학습할 수 있는 인공 신경망을 제안하였다. 운률의 세 가지 요소는 피치, 지속시간, 크기 변화가 있는데, 인공 신경망은 문장이 입력되면, 각 해당 음소의 지속시간에 따른 피치 변화와 크기 변화를 학습할 수 있도록 설계하였다.

신경망을 훈련시키기 위해 고립 단어군과 음소균형 문장군을 화자로 하여금 발성하게 하여, 녹음하고, 분석하여 운률 데이터베이스를 구축하였다.

자연음의 각 음소에 대해 지속시간과 피치변화 그리고 크기 변화를 구하여 곡선 적응 방법을 이용하여 각 변화 곡선에 대한 계수를 구해 데이터베이스를 구축한다. 이렇게 구축한 데이터베이스를 이용해 인공 신경망을 훈련시켜 평가한 결과 훈련용 데이터를 계속 확장하면 좀 더 자연스러운 운률을 발생시킬 수 있음을 관찰하였다.

I. 서 론

인간이 가장 자연스럽게 통신할 수 있는 통신 수단중의 하나가 음성이다. 정보화 시대에 들어선 지금 컴퓨터와 인터넷의 사용이 늘어나고 있으나, 아직도 컴퓨터의 기본 입출력 수단은 키보드와 모니터이다. 컴퓨터와 인간 사이의 보다 편리한 통신 매체인 음성을 이용하기 위해 음성 인식과 합성에 대한 연구가 필요했던 것이다.

1970년대부터 컴퓨터와 디지털 신호처리기술이 발달하여 디지털 음성합성에 대한 연구가 본격적으로 시작되었다. 문-음성 합성기의 합성음의 이해도와 자연감을 증가시키기 위해서는 문장내의 각음소에 대한 정확한 음향-음성학적 정보를 찾아내, 표현해주어야 한다. 그러나 대부분의 문-음성 합성기는 언어학적 정보나 자연음으로부터 추출한 정보를 바탕으로 작성한 운률법칙을 합성기에 이용하고 있다. 그러나 구현된 운률법칙이 부정확하거나 불충하고 또는 잘못 만들어진 운률법칙을 적용하게되면 합성음의 음질은 떨어질 수밖에 없다.

이러한 문제를 해결하는 방법으로 문장내의 운률을 학습할 수 있는 인공 신경망을 제안하였다. 문장 내의 각 음소의 피치와 크기 변화를 발생시키는 신경망을 각각 설계하여 학습시킬 수 있도록 설계하였다. 신경망을 훈련시키기 위해 고립 단어군과 음소 균형 문장 군으로 구성된 언어자료를 만들고 이 언어자료를 일정 환경에서 남성 화자 1인으로 하여금 3회 반복 발음하게 하여 녹음하여 음성 자료를 만들었다.

작성된 음성자료를 단기 분석하여 신경망에 필요한 훈련자료를 구축하였다. 분석에 의해 구한 각 음소의 피치 변화와 크기 변화를 2차 다항식으로 근사하는 곡선 정합 방법에 의해 각 곡선의 다항식 계수를 추출하여 신경망을 훈련시키는 자료로 만들었다. 2장에서는 한국어의 운률에 대한 고찰과 언어자료구축에 관해 논하였고, 3장에서는 운률 법칙을 훈련하기위한 신경망에 대해 논했다. 4장에서 실험 방법과 그 결과에 대해 기술하였다.

II. 한국어 문장 단위 운률

한국어 문장 단위 운률의 3가지 요소는 각 분절의 지속시간, 피치 변화, 크기 변화로 이루어지며 이들 각 분절의 운률 정보는 각 분절 고유의 특징을 포함하기도하나 다양한 주변 요인에 의해 변하게 된다. 특히 주변 분절에 의한 초분절적인 영향에 대한 연구가 많은 실정이다. 운률에 영향을 주는 요인으로는 이외에도 화자의 개성이나 감정 상태 등 다양한 변화 요인이 있을 수 있다. 이 모든 변화 요인을 다 고려할 수 는 없기 때문에 화자의 개인적인 특징이 발성단계에서 개입되지 않도록 제한해주도록 한다.

운률에 영향을 주는 요소로 먼저 의미론적 요소인 대비, 강조, 화자의 감정 상태와 발음 속도 등이 있는데 이들 모두를 감안한 언어자료를 구축하는 것은 어려우므로 평정한 상태에서 발성하는 것으로 제한한다. 구문론적인 측면에서는 실제 대화체 문장의 발음을 이용해 모델링할 수는 없고 고립 단어와 평서문에서 구문의 구, 절 등의 경계와 단어의 강세 유형 그리고 분절에 의한 영향을 고려한 운률 법칙을 모델링하였다.

이러한 점을 감안해 본 연구에서는 음소 균형 고립단어군과 문장 군을 구축하여 신경망 훈련을 위한 언어자료로 구축하였다.

구축된 언어자료를 기반으로 무향실에서 특정 남성화자 1인이 단어와 문장을 3회 반복 발음하게 하고 녹음하여 음성 자료를 만들었다.

음성자료를 단기 분석하여 각 프레임 별 10차 선형예측계수와 피치, 에너지를 구했고, 각 음소별로 구분하여 각 음소별 총 프레임 수, 피치 변화, 에너지 변화를 구해 운률에 대한 기초 자료로 만들었다.

이들 지속시간과 피치 변화, 에너지 변화를 2차 다항식으로 근사하기 위해 곡선 정합 방법을 적용하여 신경망 훈련을 위한 운률 자료를 구축하였다.

III. 피치와 에너지 곡선 발생용 인공 신경망

문-음성 합성기의 합성음질은 이해도와 자연감으로 평가하게 되는데, 기존의 합성기는 이해도는 일정 수준까지 끌어올리고 있으나, 자연감 문제는 아직 합성음이 기계음처럼 들리고 있고, 연결합성 방식에서는 자연감이 크게 늘어났으나 아직도 연결 부위의 부자연스러운 변화에 의해 자연음에는 못미치고 있는 실정이다. 이같이 합성음의 자연감이 떨어지는 요인은 연결합성이나 법칙합성 시 구현되는 운률법칙이 부정확한 때문이다.

이와 같이 운률법칙을 정확히 표현할 수 없을 때, 인공 신경망으로 하여금 문장 내에 내재하고 있는 운률 법칙을 학습하도록 하면 알고리듬화 하기 어려운 부분도 신경망이 학습하여 구현할 수 있다. 또한 훈련 자료를 계속 늘려가면 모든 가능한 경우의 운율 법칙을 학습시킬 수 있을 것이다.

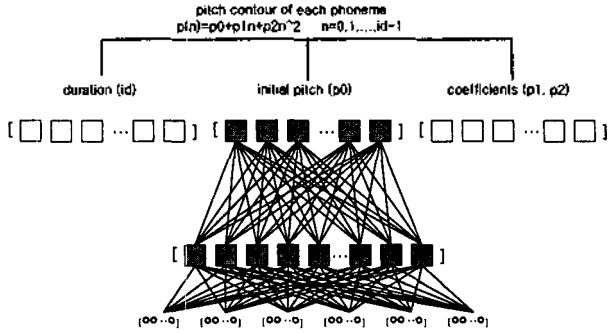

그림 1. 피치 발생 역전파 신경망의 구조
Fig.1 Architecture of BP network

피치 변화와 에너지 변화를 학습하는 인공 신경망으로 역전파 신경망을 사용하였다. 그림 1. 은 피치 변화를 학습하여 발생시키는 인공 역전파 신경망의 구조이다. 에너지 변화 신경망도 동일한 구조를 갖는다. 신경망의 입력은 문장의 음소열이 사용된다. 은닉층은 한층 사용하고, 출력층은 피치 신경망의 경우 해당 음소의 피치 변화 곡선의 다항식 계수를 내보낸다.

한국어 문장의 경우 음운 변화를 거치면 각 음절에 초성 자음 18가지와, 중성 모음 21가지, 종성 자음 7가지가 남게된다. 이들 음소 이외에도 쉼표나 마침표 등의 구문 부호가 포함되므로 입력 문장의 각 음소를 표현하기 위해 필요한 비트 수는 8 비트를 지정하였다. 나중에 필요하면 다양한 구분 부호를 추가할 수 있을 것이다.

한국어 문장의 경우 문장 내에 몇 개의 운율 구가 존재하는 것으로 연구 조사되었다. 이러한 운률구의 경계에 대한 정보도 입력 단에 포함된다면 신경망을 더 효율적으로 학습시킬 수 있을 것이다.

한 운률구 내의 음소 분절이 2개에서 10개까지 다양하므로 초분절적인 요인을 감안한다면 적어도 신경망의 입력 단의 노드 수를 11개 이상으로 지정해야 할 것이다. 각 노드에 8 비트를 할당하였으므로 입력 층의 총 비트 수는 88 비트가 된다.

이 11개의 음소열 중 6번째 음소의 운률 정보를 출력층에 목표 패턴으로 제시하여 신경망을 학습시킨다. .

신경망의 비선형 사상을 위해 1개의 은닉층을 사용하였고 은닉층의 노드의 수는 입력층의 노드 수오 같게 지정하였다.

출력층은 입력층의 중앙 음소에 대한 2차 다항식 계수를 출력한다. 다항식 계수와 초기치, 지속시간을 출력하므로 4개의 모듈로 구성된다.

각 계수와 초기치, 지속시간에 각각 16비트씩을 할당하였다.

10KHz로 표본화한 음성 자료를 단기 분석하면

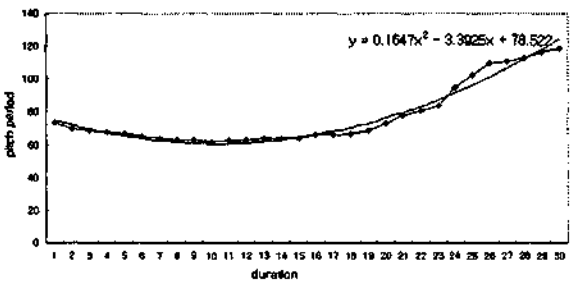

Fig. 2 Pitch contour of '예' phoneme and approximant line

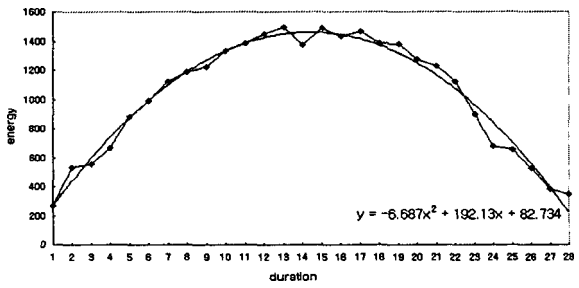

Fig. 3 Energy contour of '예' phoneme and approximant line

각 문장의 피치와 에너지 변화 곡선을 그릴 수 있다. 각 프레임의 표본 수를 256 표본으로 하고 128 표본씩 이동시켜 운률 정보를 계산하였다. 이 운률 곡선과 선형예측 계수 변화 곡선을 이용하여 각 음소로 구분하였다. 각 음소의 지속시간이 1프레임에서 24 프레임까지 변화하는 것을 알 수 있었다. 이 중에서 피치 변화 곡선과 에너지 변화 곡선을 다항식으로 근사할 수 있는데, 여기서는 다항식의 차수를 2차로 제한하였고, 근사 방식은 비선형 곡선 정합 방법을 사용하였다.

각 음소의 피치와 에너지 변화 곡선에 대한 2차 다항식 근사식은 다음과 같다.

$$p(n) = p2*n**2 + p1*n + p0,\ 0<=n<= d-1 \quad (1)$$

$$e(n) = e2*n**2 + e1*n + e0,\ 0<=n<= d-1 \quad (2)$$

여기서 p1, p2는 피치 계수, p0는 피치 초기치, d는 지속시간(프레임 수)이다. e1, e2는 에너지 계수이고, e0는 에너지 초기치이다.

그림 2는 음성 파형 중 '예'라는 음소의 피치변화 곡선과 곡선정합방법에 의해 구한 그 추세선을 표시한 것이다. 그림 3은 같은 음소의 에너지 변화곡선과 그 추세선을 표시한 것이다.

각 음소의 운률 정보를 신경망에 학습시키기 위해 앞에서 언급했듯이 먼저 언어 자료를 구축하고, 그 언어자료를 특정화자로 하여금 일정 조건하에서 반복 발성하게 하고, 녹음하여 음성 자료를 채록한다. 채록된 음성 자료를 단기 분석하여 신경망에 필요한 운율 자료를 구축한다.

IV. 실험

신경망 훈련을 위해 음소 균형 412개의 고립단어를 토대로 100개의 의미 문장을 구성하여 언어자료를 만들었다. 남성화자 1인이 이들 언어자료를 3회 연속 발음하도록 하고 녹음하여 음성 자료를 채록하였다. 단기 분석 기법을 사용하여 10차 선형 예측계수와 운률 정보를 도시하고 이를 근거로 각 음소 분할을 하고, 각 음소의 운률 변화 곡선을 대상으로 비선형 곡선 정합 방법에 의해 구한 운율 자료를 구했다.

신경망 훈련 단계에서는 3회 발성된 자료중 처음 2개의 자료를 이용해 입력 층에 문장을 인가하고 중앙 음소의 운률 정보를 출력 층에 목표 패턴으로 인가하여 신경망이 훈련하도록 하였다. 훈련 주기는 200회로 제한하고 그 전에 훈련을 마칠 수 있는 최소 오차 임계치를 설정하였다.

훈련 단계에서 각 신경망이 보여준 추정율은 다음과 같다.

표. 1 훈련단계의 추정율
Table 1. Estimation rates in training phase

피치 신경망

d	p0	p1	p2
92.4	91.7	92.5	92.2

에너지 신경망

d	e0	e1	e2
90.4	91.7	91.3	90.3

평가 단계에서는 평가용 자료로 3번째 자료를 사용하였다. 입력 단에 문장을 인가하고 신경망의 출력치를 3번째 자료에 있는 해당 음소의 3피치 및 에너지에 대한 다항식 계수와 비교하여 추정율을 계산하였다.

표.2 평가단계의 추정율
Table 2. Estimation rates in test phase

피치 신경망

d	p0	p1	p2
90.3	90.3	89.4	90.7

에너지 신경망

d	e0	e1	e2
88.4	90.1	89.4	88.3

V. 결론

피치 신경망의 추정율이 훈련 단계에서는 92%이고 평가 단계에서는 90%였다. 에너지 신경망의 경우 훈련단계에서는 90%, 평가 단계에서는 89%은 성능을 나타냈다. 추정율을 높이기 위해서는 우선 언어자료를 좀더 광범위하게 구축해야 하고, 입력단의 음소 수가 11개로 제한되어 있기 때문에 이 것을 벗어나는 초분절적인 요인은 아직 반영할 수 없다는 문제가 있다. 근사식의 다항식의 차수가 2차로 제한되어 있어 변화 곡선을 정확히 근사하는데 한계가 있다. 이러한 문제를 해결하기 위해서는 입력과 출력 노드 수를 늘리면 가능하겠으나 계산량이 기하급수적으로 늘어나는 문제가 있다.

참고문헌

[1] J. D. Markel and A. H. Gray Jr., *Linear Prediction of Speech.* Springer- Verlag, 1976.

[2] J. Allen, M. S. Hunnicutt and D. H. Klatt et al, *From Text To Speech.* Cambridge University Press, 1987.

[3] A. Waibel, *Prosody and Speech Recognition.* Morgan Kaufmann Publishers, 1988.

[4] A. M. Liberman et al, "Minimal rules for synthesizing speech," J. Acoust. Soc. Am., vol.31, No.11, pp.1490-1499, Nov. 1959.

[5] J. Allen, "Synthesis of speech from unrestricted text," Proc. IEEE, vol.64, No.4, pp.433-442, Apr. 1976.

[6] N. Umeda, "Vowel duration in American English," J. Acoust. Soc. Am., vol.56, pp.434-445, 1975.

[7] J. Pierrehumbert, "Synthesizing intonation," J. Acoust. Soc. Am., vol.70, No.4, pp.985-995, Oct. 1981.

[8] R. M. Meli and F. Fallside, "The modeling of F0 contours," in IEEE Proc. ICASSP'82, 1982, pp.947-949.

[9] Hyun Bok Lee, "Korean prosody : Speech rhythm and intonation," Korea Journal, pp.42-69, Feb. 1987.

[10] M. Ljungqvist and H. Fujisaki, "Generating Intonation for Swedish Text-to-Speech Conversion Using a Quantitative Model for the F0 Contour," in Proc. Eurospeech '93, 1993, pp.873-876.

[11] C. Tuerk and T. Robinson, "Speech Synthesis Using Artificial Neural Networks Trained on Cepstral Coefficients," in Proc. Eurospeech '93, 1993, pp.1713-1716

[12] J. C. Lee, S. H. Kim and M. Hahn, "Intonation Processing for Korean TTS Conversion Using Stylization Method," in Proc. ICSPAT '95, 1995, vol.II, pp.1943-1946.

[13] M. Riedi, "A Neural-Network-Based Model of Segmental Duration for Speech Synthesis," in Proc. EUROSPEECH '95, 1996, vol.I, pp.599-602.

[14] D. P. Morgan and C. L. Scofield, *Neural Networks and Speech Processing,* Kluwer Academic Pub., 1991.

[15] Mazin G. Rahim, *Artificial Neural Networks for Speech Analysis/Synthesis,* Chapman & Hall, 1994.

[16] Adam Blum, *Neural Networks in C++,* John Wiley & Sons Inc., 1992.

[17] Sok Wang Chang, Hyun Joon Kim, Chang Su Ryoo, Un Cheon Lim, "A Study on the Prosody Generation in Isolated Words with an Artificial Neural Network," in Proc. ICSP'97, 1997, Vol. 1 of 2, pp. 207 - 210

[18] Hyun Joon Kim, Chang Su Ryoo, Sok Wang Chang, Un Cheon Lim, " A Study on the Prosodic Marker in a Korean Sentence," in Proc. ICSP'97, 1997, Vol. 1 of 2, pp. 213-216.

[19] Bong Wan Kim, Sun Tae Kim, Tae Wan Kim, Young Il Lee, Yong Ju Lee, " Design and Construction of Korean Speech Database for Common Use," in Proc. ICSP'97, 1997, Vol. 1 of 2, pp. 759-762.