한국어 운율 발생용 인공신경망의 구조 및 학습에 관한 연구

민경중, 임운천
호서대학교 대학원 전자공학과

A Study on the Architecture and Learning of the Artificial Neural Networks for Prosody Generation of Korean Sentence

Kyung-Joong Min, Un-Cheon Lim
Dept. of Electronic Eng., Graduate School, Hoseo University
uclim@office.hoseo.ac.kr

요약

음성처리기술은 정보화 시대를 위한 주요 기술의 하나이다. 이 중에서도 음성합성의 연구는 디지털 신호처리 기술과 컴퓨터의 발달로 활발히 진행되고 있다. 그러나 음성 합성기에 의해 발생된 합성음의 음질은 이해도 면에서는 상당한 진전이 있었지만, 자연감 면에서는 만족한 수준에 도달할 수 없었는데, 이러한 합성시스템의 문제점을 해결하는 방법은 다양하게 적용되는 언어정보와 합성음의 자연감을 결정하는 정확한 운율정보가 필요하다. 그러나 구한 운율 정보가 자연음에 존재하는 모든 운율 법칙을 포함할 수 없고, 또한 추출한 운율 법칙이 틀린 것이면 자연감이나 이해도가 떨어지는 합성음이 만들어지고, 이것은 음성 합성 시스템의 실용화에 장애로 작용할 것이다.

본 논문은 한국어 음성 합성 시 문제가 되는 자연감을 높이기 위한 한 방법으로 자연음에 내재하는 운율 변화를 효율적으로 학습할 수 있는 인공 신경망을 제안하였다.

I. 서 론

인간의 가장 기본적인 통신수단 중의 하나인 음성을 컴퓨터와 인간사이 통신매체로 사용하기 위해서는 컴퓨터에 의한 음성인식과 합성이 필요하다. 최근에 멀티미디어나 이동통신 등과 같은 다양한 매체를 통한 정보통신에서도 음성이 가장 자연스러운 정보교환 수단이 되고 있다.

문-음성 합성기의 합성음의 이해도와 자연감을 증가시키기 위해서는 문장 내의 각 음소에 대한 정확한 음향-음성학적 정보를 찾아내 표현해 주어야 한다. 대부분의 문-음성 합성기는 언어학적 정보나 자연음에서 추출한 운율 정보를 바탕으로 작성한 운율 법칙을 합성기에 이용하고 있다. 다양한 운율 법칙을 구하기 위해서는 많은 자연음 데이터가 필요할 것이나, 여기서 구한 운율 법칙이 부정확하거나 불충분하고 또는 잘못 만들어지면 합성음의 음질은 떨어질 수밖에 없다.

이러한 문제를 해결하는 방법으로 한국어 문장을 발성한 자료를 대상으로 문장 내에 존재하는 운율을 학습하여 출력할 수 있는 인공 신경망을 제안하였다. 이를 위해 문장 내 각 음소의 지속시간에 따른 피치와 크기 변화를 선형예측분석을 통하여 구하여 인공신경망을 훈련 및 측정하는데 사용하였다. 문장 내 각 음소의 피치와 에너지 변화 곡선을 구하고, 이것을 2차 다항식으로 근사하여 구한 다항식 계수를 인공 신경망의 목표 패턴으로 제시하여 훈련시킨 인공 신경망은 문장의 음소열을 입력패턴으로 제시하면 문장 내의 중앙에 해당하는 음소의 운율 정보 즉 지속 시간과 피치 및 에너지를 학습하여 출력하도록 설계하였다.

II. 한국어 문장 단위 운율

일반 발성 언어에서 운율의 구성 요소로는 각 분절의 지속시간, 피치 그리고 에너지 변화가 있다. 문장 단위 음성에서는 억양, 악센트, 리듬, 지속시간 등의 운율로 나타낼 수 있다.

일반적으로 문장 단위의 운율에 대한 연구는 주로 피치와 지속 시간에 대해 이루어졌다. 그리고 운율의 한 요소인 에너지의 변화에 대한 연구는 극히 미미한 실정인데, 특히 음성 합성 중 연결 합성 방식에서는 에너지의 급격한 변화가 합성음의 자연감을 저하시키는 주된 요인이 될 수 있다.

본 논문에서는 의미론적으로는 평정한 상태에서 발성하는 것으로 제한하고, 구문론적으로는 음소 균형고립 단어 군을 기반으로 작성한 구, 절 단위 문장 군을 평탄하게 발성할 수 있는 언어 자료를 구축하여 문장 내의 구, 절 등의 경계와 단어의 강세 유형 그리고 분절에 의한 영향을 고려한 운율 법칙을 학습할 수 있게 하였다.

구축된 언어 자료를 기반으로 음성 시료를 만들어, 단기 분석하여 각 프레임 별 10차 선형 예측 계수와 피치 및 에너지를 구하여 각 음소별 총 프레임 수(지속 시간)와 피치 및 에너지 변화를 구했다.

이들 지속 시간에 따른 피치와 에너지 변화 곡선을 다항식으로 근사하여 지속 시간과 다항식 계수로 이루어진 운율 패턴을 구해, 인공 신경망을 학습시키고 평가할 수 있는 목표 및 평가패턴으로 사용할 수 있다.

본 논문에서는 피치 변화곡선과 에너지 변화곡선을 근사하는 다항식의 차수를 2차로 하여 실험하였고, 근사 방식은 비선형 곡선 정합 방법을 사용하였다.

각 음소의 피치와 에너지 변화곡선을 근사하는 2차 다항식은 다음과 같다.

$$p(n) = p_2 n^2 + p_1 n + p_0, \qquad 0 \le n \le d-1 \quad (1)$$

$$e(n) = e_2 n^2 + e_1 n + e_0, \qquad 0 \le n \le d-1 \quad (2)$$

여기서 p_1, p_2는 피치 계수, p_0은 피치 초기 치, d는 지속시간(프레임 수)이다. e_1, e_2는 에너지 계수이고, e_0는 에너지 초기 치이다.

단기분석에 의해 구해진 피치 및 에너지 변화곡선에 대해 곡선정합방법을 적용하여 지속시간과 다항식계수를 구해 이를 인공신경망 운율발생기를 훈련 및 평가하기 위한 운율자료로 사용하였다.

III. 인공 신경망의 설계 및 훈련

인공 신경망은 계층적 구조인 입력 층, 은닉 층, 출력 층의 3개의 층으로 구성하였다. 문장의 음소 열 조합이 입력 층에 인가되면 인공 신경망은 이 음소열의 중앙에 해당하는 음소의 운률을 학습하여 자연음의 운율과 유사한 합성운율을 발생시킨다. 입력 층은 11개의 음소 열을 나타내는 유니트로 구성하였다. 출력 층은 입력 음소열의 중앙에 해당하는 음소의 피치와 에너지 계수를 학습하여 출력하도록 설계하였다.

다음 그림 1은 실험에 사용한 역전파 인공 신경망의 구조를 나타낸 것이다.

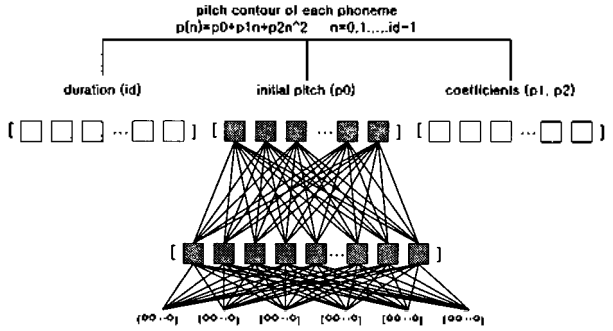

그림 1. 역전파 인공 신경망의 구조

Fig.1 Architecture of Back Propagation network

III-1 입력 층

한국어의 경우 일반 문장에 음운 변동을 적용한 후 이를 통해 구한 대표 음소는 자음 18개, 중성 모음 21개, 종성 자음 7개가 남게 된다. 이들 대표 음소와 경계 구간을 나타내는 마침표와 쉼표, 그리고 blank(□)도 하나의 음소로 포함시켜 입력 단에 인가할 수 있다. 이들 대표 음소와 심볼들을 표현하기 위해 6 bit를 할당하고, 운율구 내에서의 음소의 상대적인 위치에 관한 정보와 총 음소 수에 대한 정보를 위해 각각 5 bit를 할당하여, 각 유니트에 총 16비트를 할당하여 인공 신경망의 입력 단을 구성하였다[11].

III-2. 출력 층

각 음소의 피치와 에너지 변화는 회귀분석을 통한 추세선을 이용하여 2차 다항식으로 근사화한 후 그 계수와 초기 피치(에너지)를 구하였다. 그리고 다항식의 변수인 지속 시간은 그 프레임 수를 출력 패턴의 일부로 사용하였다. 무성자음의 경우 피치가 존재하지 않으므로 피치 변화 곡선의 계수와 초기 피치 값은 모두 0으로 지정하였다.

문-음성 합성기의 합성음질은 이해도와 자연감으로 평가하게 되는데, 기존의 합성기에서 이해도는 일정 수준까지 끌어올리고 있으나, 연결 합성방식에서는 자연감이 크게 늘어났으나 아직도 연결 부위의 부자연스러운 변화를 감지할 수 있는 실정이다. 이같이 합성음의 자연감이 떨어지는 요인은 연결합성이나 법칙합성 시 구현되는 운율법칙이 부정확한 때문이다.

이와 같이 운율법칙을 정확히 표현할 수 없을 때, 인공신경망으로 하여금 문장 내에 내재하고 있는 운율법칙을 학습하도록 하면 알고리듬으로 구현하기 어려운

부분도 인공신경망이 학습하여 구현할 수 있다.

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

한 운율 구 내의 음소분절이 2개에서 10개까지 다양하므로 초 분절적인 요인을 감안한다면 적어도 인공신경망의 입력 단의 노드 수를 11개 이상으로 지정해야 하며, 11개의 음소열 중 중앙에 해당하는 6번째 음소의 운율정보를 출력 층에 목표패턴으로 제시하여 인공신경망을 학습시킨다. 중앙 음소 전/후 5개 음소에 의한 초 분절적 영향을 그대로 학습할 수 있고, 일정 임계치 이내로 허용오차가 떨어지면 음소 열을 전진 이동시켜 다음 음소가 중앙 음소가 되게 하여 훈련과정을 반복한다.

인공 신경망의 비선형 사상을 위해 1개의 은닉 층을 사용하였고 은닉 층의 노드의 수는 입력 층의 노드 수와

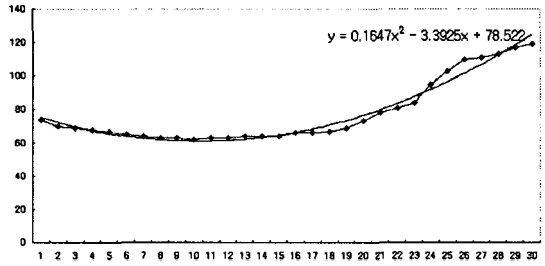

그림 2. 모음 /에/의 피치변화곡선과 그 근사선

Fig.2 The pitch contour of Korean vowel /ye/ and it's approximated line

같게 지정하였다.

출력 층은 입력 층의 중앙음소에 대한 2차 다항식 계수와 초기 치, 지속 시간을 출력하므로 4개의 모듈로 구성된다.

이러한 신경망의 학습을 통한 정확한 운율법칙을 구현하기 위해 2차 다항식 계수를 구하였다.

위 그림 2는 모음 /에/의 피치변화곡선과 그 근사선을 표시한 것이다. 이 근사식의 계수를 인공 신경망의 출력패턴으로 사용하게 된다.

IV. 실험

인공신경망 훈련을 위해 음소 균형 412개의 고립단어를 토대로 100개의 의미문장을 구성하여 언어자료를 만들었다. 남성화자 1인이 이들 언어자료를 3회 연속 발음하도록 하고 녹음하여 음성 자료를 채록하였다.

10 KHz로 표본화한 음성 자료를 단기 분석하면 각 문장의 피치와 에너지 변화 곡선을 그릴 수 있다. 각 프레임의 표본 수를 256 표본으로 하고 128 표본씩 이동시켜 운율정보를 계산하였다. 이 운율 곡선과 선형 예측계수 변화 곡선을 이용하여 각 음소로 분할하였다.

각 음소의 지속시간이 1 프레임에서 24 프레임까지 변화하는 것을 알 수 있었다. 이를 근거로 음소를 분할하고, 이들 음소의 운율변화 곡선을 대상으로 비선형 곡선 정합 방법에 의해 지속 시간과 다항식 계수를 구해 운율 자료를 구축하였다.

인공 신경망의 훈련 단계에서는 3회 발성된 음성 자료에서 구한 운율 자료 중 처음 2개의 운율 자료를 이용해 인공 신경망을 훈련시켰다.

훈련 주기는 200회로 제한하고 그 전에 훈련을 마칠 수 있는 최소 오차 임계치를 설정하였다. 훈련 단계에서 각 인공 신경망이 보여준 추정율은 다음 표 1과 같다.

표. 1 훈련 단계의 추정율

Table 1. Estimation rates in training phase

피치 신경망의 2차 다항식 추정율

d	p_0	p_1	p_2
92.7	92.0	92.9	92.6

에너지 신경망의 2차 다항식 추정율

d	e_0	e_1	e_2
90.8	91.9	91.6	90.7

평가단계에서는 학습을 마친 인공신경망에 3번째 운율자료를 평가용 자료로 제시하여 인공신경망의 실제 출력 패턴과 비교하여 인공 신경망의 추정율을 계산하여 그 결과를 표 2에 나타내었다.

표.2 평가단계의 추정율

Table 2. Estimation rates in test phase

피치 신경망의 2차 다항식 추정율

d	p_0	p_1	p_2
90.8	90.8	89.8	91.1

에너지 신경망의 2차 다항식 추정율

d	e_0	e_1	e_2
88.9	90.5	89.8	88.8

V. 결론

피치 신경망의 추정율이 훈련 단계에서는 92.5%이고 평가 단계에서는 90.6%이었다. 에너지 신경망의 경우 훈련 단계에서는 91.3%, 평가 단계에서는 89.5%의 성능을 나타냈다.

추정율을 높이기 위해서는 우선 언어 자료를 좀 더 광범위하게 구축하고, 발성 횟수를 늘려 훈련량을 늘리는 방법이 있겠으나 그 처리량이 늘어나는 문제가 있다. 반면 훈련용 자료가 적으면 인공 신경망이 해당 자료를 과도 학습하는 문제가 발생하게 되므로 적절한 규모의 언어 및 음성, 운률 자료의 구축을 해줄 필요가 있을 것이다.

운율변화는 주로 운율 구를 중심으로 나타나므로, 본 논문에서는 입력 단의 음소 수를 11개로 제한하였으므로, 해당 음소 전후 5음절 범위 이내의 초 분절적인 영향은 잘 반영할 수 있으나, 그 이상의 문장 전체에 걸친 변화는 학습할 수 없는 문제가 있다. 이것은 입력단 음소열의 수를 증가시키면 해결할 수 있을 것이나 계산량이 늘어나는 문제가 있다.을 것이다.

참고문헌

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

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

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

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

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

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

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

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

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

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

[11] Kyung-Joong Min, Joon-Sik Kim. Un-Cheon Lim, "Input/Output Pattern of Neural Networks for Prosody Generation of Korean Sentences," in Proc. ICSP'99, 1999, Vol. 1 of 2, pp. 161-166.

[12] 신동엽, 민경중, 강찬구, 임운천, " 한국어 운율발생용 인공 신경망의 입출력 패턴에 관한 연구," 제17회 음성통신 및 신호처리 학술대회 논문집, 제17권 제1호, pp. 245-248.

[13] 신동엽, 임운천, "한국어 운율 발생을 위한 인공 신경망의 구조에 관한 연구," 2001년도 한국음향학회 학술발표대회논문집, 제20권 제1(s)호, pp. 307-310.

[14] 민경중, 임운천, "인공 신경망의 한국어 운율 발생에 관한 연구," 2001년도 한국음향학회 학술발표대회 논문집, 제20권 제1(s)호, pp. 311-314.

[15] Dong-Yup Shin, Chan-Goo Kang, Un-Cheon Lim, " Prosody Generation of Artificial Neural Networks in Korean Sentences," *Proc. of ICSP 2001*, 2001, Vol. 2 of 2, pp. 771-776.

[16] Kyung-Joong Min, Un-Cheon Lim, " Architecture of Artificial Neural Networks for Prosody Generation in Korean Sentences," *Proc. of ICSP 2001*, 2001, Vol. 2 of 2, pp. 819-823.

[17] 김순효, 민경중, 임운천, " 인공신경망 운율발생기의 한국어 운율학습에 관한 연구," 제19회 음성통신 및 신호처리 학술대회 논문집, 제19권 제1호, pp. 133-136.

[18] 김순효, 민경중, 강찬구, 임운천, " 한국어 운율 발생용 인공산경망의 구조 및 설계에 관한 연구," 제20회 음성통신 및 신호처리 학술대회 논문집, 제20권 제1호, pp. 305-308.

[19] 민경중, 임운천, "운률 발생 인공신경망 설계 및 학습," 2003년도 한국음향학회 학술발표대회 논문집, 제22권 제1(s)호, pp. 145-148.

[20] Kyung-Joong Min, Un-Cheon Lim, "Korean Prosody Generation and Artificial Neural Networks," *Proc. of INTERSPEECH 2004-ICSLP*, 2004.

[21] Kyung-Joong Min, Chan-Goo Kang, Un-Cheon Lim, "Number of Output Nodes of Artificial Neural Networks for Korean Prosody Generation," *Proc. of INTERSPEECH 2004-ICSLP*, 2004.