인공 신경망의 한국어 운율 학습

신동엽, 민경중, 임운천
호서대학교 대학원 전자공학과

Learning of Artificial Neural Networks about the Prosody of Korean Sentences.

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

요 약

음성 합성기의 합성음의 자연감을 높이기 위해 자연음에 내재하는 정확한 운율 법칙을 구하여 음성합성 시스템에서 이를 구현해 주어야 한다. 무제한 어휘 음성합성 시스템의 문-음성 합성기에서 필요한 운율 법칙은 언어학적 정보를 이용해 구하거나, 자연음에서 추출하고 있다. 그러나 추출한 운율 법칙이 자연음에 내재하는 모든 운율 법칙을 반영하지 못했거나, 잘못 구현되는 경우에는 합성음의 자연성이 떨어지게 된다.

이런 점을 고려하여 본 논문에서는 한국어 자연음을 분석하여 추출한 운율 정보를 인공 신경망이 학습하도록 하고 훈련을 마친 인공 신경망에 문장을 입력하고, 출력으로 나오는 운율 정보와 자연음의 운율 정보를 비교한 결과 제안한 인공 신경망이 자연음에 내재하고 있는 운율을 학습할 수 있음을 알 수 있었다.

운율의 3대 요소는 피치, 지속시간, 크기의 변화이다. 제안한 인공 신경망이 한국어 문장의 음소 열을 입력으로 받아들이고, 각 음소의 지속시간에 따른 피치 변화와 크기 변화를 출력으로 내보내면, 자연음을 분석해 구한 각 음소의 운율 정보인 목표 패턴과 출력 패턴의 오차를 최소화하도록 인공 신경망의 가중치를 조절할 수 있도록 설계하였다.

지속시간에 따른 각 음소의 피치와 크기 변화를 학습시키기 위해 피치 및 크기 인공 신경망을 구성하였다. 이들 인공 신경망을 훈련시키기 위해 먼저 음소 균형 문장 군을 구축하여야 하고, 이들 언어 자료를 특정 화자가 일정 환경에서 읽고 이를 녹음하여, 분석하여 구한 운율 정보를 운율 데이터베이스로 구축하였다.

문장 내의 각 음소에 대해 지속 시간과 피치 변화 그리고 크기 변화를 구하고, 곡선 적응 방법을 이용하여 각 변화 곡선에 대한 다항식 계수와 초기 값을 구해 운율 데이터베이스를 구축한다. 이 운율 데이터베이스의 일부는 인공 신경망을 훈련시키는데 이용하고, 나머지로 인공 신경망의 성능을 평가하여 인공 신경망이 운율 법칙을 학습할 수 있었다.

언어 자료의 문장 수를 늘리고 발음 횟수를 늘려 운율 데이터베이스를 확장하면 인공 신경망의 성능을 높일 수 있고, 문장 내의 음소의 수를 감안하여 인공 신경망의 입력 단자의 수는 계산량과 초분절 요인을 감안하여 결정해야 할 것이다.

1. 서 론

인간의 가장 자연스러운 통신 수단의 하나인 음성을 기계와 인간 사이의 통신 수단으로 사용하기 위한 시도는 오래 전부터 있었고 현재도 지속적으로 연구가 되고 있다.

컴퓨터로 가능한 음성 합성시스템인 문-음성 합성기의 합성음의 이해도와 자연감을 증가시키기 위해서는 문장 내의 각 음소에 대한 정확한 음향-음성학적 정보를 찾아내어 합성해주어야 한다. 대부분의 문-음성 합성기는 언어학적 정보나 자연음을 분석하여 구한 정보를 바탕으로 추출한 운율법칙을 합성기에 이용하고 있다. 그러나 구현된 운율법칙이 부정확하거나 불충분하고 또는 잘못 만들어진 운율법칙을 적용하게 되면 합성음의 음질은 떨어지게 된다.

이러한 문제를 해결하는 방법으로 문장 내의 운율 법칙을 학습할 수 있는 인공 신경망을 제안하고 문장 구조와 인공 신경망의 운율 법칙 학습에 대해 고찰하였다.

인공 신경망을 훈련시키기 위해 음소 균형 문장 군으로

구성된 언어 자료를 구축하고, 이 언어 자료를 일정 환경에서 남성 화자 1인으로 하여금 3회 반복 발음하게 하여 녹음하고, 음성 시료를 채록하였다. 작성된 음성 시료를 대상으로 단기 분석을 행하여 각 음소에 대한 원시 운율 자료를 구했다.

곡선 정합 방법을 이용해 원시 운율 자료 내의 각 음소의 피치 변화와 크기 변화 곡선을 2차 다항식으로 근사하여 각 변화 곡선의 다항식 계수와 초기치를 구해 인공 신경망을 훈련시키고, 평가할 수 있는 운율 자료로 만들었다. 문장 내 각 음소의 운율 정보를 학습하기 위해 피치 및 에너지(크기) 인공 신경망을 구축하여 훈련을 통해 이들 신경망이 자연음의 운율 변화를 학습할 수 있도록 하였고 일정 횟수 동안 훈련을 시킨 다음 그 출력을 자연음의 운율과 비교하여 인공 신경망이 운율을 학습할 수 있음을 알 수 있었다.

2장에서는 한국어의 운율에 대한 고찰과 언어자료 구축에 관해 논하였고, 3장에서는 운율 법칙을 학습하기 위한 인공 신경망에 대해, 4장에서는 실험 방법과 그 결과에 대해 기술하였다.

II. 한국어 문장의 운율 변화

문장 내의 각 분절의 운율 정보는 각 분절 고유의 특징을 포함하면서 다양한 주변 요인에 의해 변하게 된다. 특히 주변 분절에 의한 초 분절적인 영향에 의해 각 분절의 운율은 변하게 된다. 구문론적인 영향 외에도 각 분절의 운율에 영향을 주는 요인으로 화자의 개성이나 감정 상태 등이 있을 수 있다.

이 모든 변화 요인을 반영한 언어 자료를 구축하고 다양한 발성 환경, 화자에 대한 음성 자료를 구해야 한다. 그러나 이렇게 하기 위해서는 언어 자료 구축, 분석, 훈련 등에 막대한 시간과 노력을 투자해야 하므로, 본 논문에서는 화자의 개인적인 특징이 발성 단계에서 개입되지 않도록 하기 위해 평정한 상태에서 문장을 발성하는 것으로 제한하였다. 또한 구문론적인 측면에서는 실제 대화체 문장의 발음을 모델링 할 수는 없기 때문에, 평서문에서 구문의 구, 절 등의 경계와 단어의 강세 유형 그리고 분절에 의한 영향을 반영한 운율 법칙을 인공 신경망이 학습하도록 제한하였고, 음성학적으로 균형 잡힌 고립단어 군을 기반으로 문장과 구를 작성하여 제한된 문장의 언어 자료를 구축하였다.

구축된 언어 자료를 기반으로 무향실에서 특정 남성 화자 1인이 언어 자료를 3회 반복하여 발음하게 하고, 이것을 녹음하여 음성 자료를 만들었다

음성 자료를 단기 분석하여 각 프레임별 10차 선형예측계수와 피치, 에너지를 구했고, 각 음소별로 분할하여 각 음소별 총 프레임 수(지속시간)와 피치 변화 그리고 에너지 변화를 구해 운율에 대한 원시 자료로 만들었다. 각 음소의 지속시간과 피치 변화, 에너지 변화를 2차 및 3차 다항식으로 근사하기 위해 곡선 정합 방법을 적용하여 초기 치와 다항식 계수를 구해 신경망 훈련과 평가를 위한 운율 자료로 구축하였다.

III. 피치와 에너지 인공 신경망

음성 합성 방식은 법칙합성과 연결합성 방식으로 크게 나눌 수 있으며 연결 합성 방식에서는 합성 단위를 다양하게 하여 운율 법칙의 정확한 구현이 어려운 분절간 천이 구간 전부를 하나의 합성 단위로 사용하여 자연감은 늘어났으나, 합성용 데이터베이스의 규모가 커지는 문제가 있다. 법칙 합성 방식에서는 음소 단위의 운율 변화 법칙을 구현하여 합성음을 생성시키기 때문에 DB의 규모는 작으나 자연감이 떨어진다. 두 방식 모두 운율 법칙을 구현하여 합성음의 자연감과 이해도를 높이려고 하고 있다.

이와 같이 구현할 운율 법칙을 정확히 표현할 수 없을 때, 인공 신경망으로 하여금 문장 내에 내재하고 있는 운율 법칙을 학습하도록 하면 법칙으로 만들기 어려운 부분도 인공 신경망이 학습하여 구현할 수 있을 것이다. 또한 훈련용 운율 자료를 계속 늘려 가면 모든 가능한 경우의 운율 법칙을 학습시킬 수 있을 것이다.

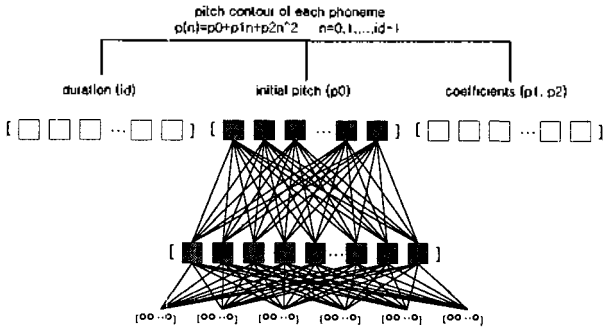

그림 1. 역전파 신경망의 구조
Fig.1 Architecture of BP network

피치와 에너지 변화를 학습하는 인공 신경망으로 역전파(Back Propagation) 신경망을 사용하였다. 그림 1에 피치 변화를 학습하여 발생시키는 역전파 신경망의 구조를 나타내었다. 에너지 신경망도 피치 신경망과 동일한 구조를 갖고 항상 동시에 동일 문장에 대해 훈련을 받도록 하였다.

인공 신경망의 입력 단에 한 문장의 음소 열을 인가하

도록 설계하였다. 1개의 은닉 층을 통고하여 출력 층에서는 피치 신경망의 경우 해당 음소의 피치 변화 곡선의 다항식 계수와 초기 치, 지속 시간이 출력되도록 설계하였다.

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

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

한 운율 구 내의 음소 분절의 수가 2개에서 10개 이상까지 변하므로 초 분절적인 요인과 계산량을 감안하여 인공 신경망의 입력 단의 노드 수를 11개로 하였다. 각 노드에 8 비트를 할당하였으므로 입력 층의 총 비트 수는 88 비트가 된다. 이 11개의 음소열 중 6번째 음소의 운율 정보를 출력 층에 목표 패턴으로 제시하여 인공 신경망을 학습시킨다.

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

10 KHz로 표본화한 음성 자료를 단기 분석하면 각 문장의 피치와 에너지 변화 곡선을 구할 수 있다. 각 프레임의 표본 수를 256 표본으로 하고 128 표본씩 이동시켜 운율 정보를 계산하였다. 단기 분석에 의해 구한 운율 곡선과 선형 예측계수 변화곡선을 이용하여 각 음소로 분할한 결과, 고립 단어에서는 각 음소의 지속시간이 1 프레임에서 24 프레임까지 변화하는 것으로 나타났으며, 일반적으로 운율 구를 기반으로 변화하므로 프레임 수는 늘어나지 않음을 알 수 있다.

이러한 점을 감안하여 출력 층은 입력된 음소열 중 중앙에 해당하는 음소에 대한 2차 다항식 계수와 초기 치, 지속시간(프레임 수)을 출력하는 4개(혹은 5개)의 모듈로 구성하고 각 다항식 계수와 초기 치에 16비트 그리고 지속시간에 8비트를 할당하였다.

각 음소의 피치 변화 곡선과 에너지 변화 곡선을 다항식으로 근사할 수 있으며 근사 방식은 비선형 곡선 정합 방법을 사용하였다.

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

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

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

여기서 식 (1)의 p_1, p_2는 피치 변화곡선의 다항식 계수, p_0는 피치 변화 곡선의 초기치, d는 음소의 지속 시간(프레임 수)이다. 식(2)의 e_1, e_2는 에너지 변화곡선의 다항식 계수이고, e_0는 에너지 변화곡선의 초기 치이고 d는 음소의 지속 시간이다.

그림 2는 음소 '예'의 피치변화 곡선과 곡선 정합방식으로 구한 2차 다항식 근사 곡선을 표시한 것이다.

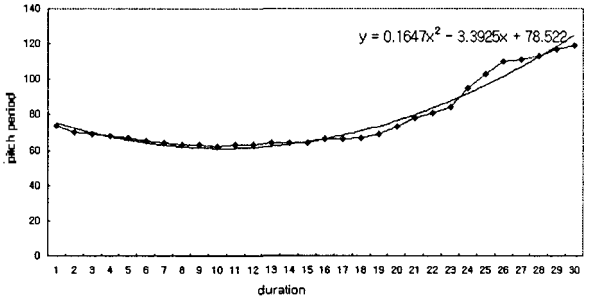

그림 2. 음소 /예/의 피치변화 곡선과 그 추세선
Fig. 2 Pitch contour of a phoneme /ye/ and it's regressive line

Ⅳ. 실 험

인공 신경망을 훈련시키고 평가를 하기 위해 음소 균형 412개의 고립 단어를 기반으로 100개의 의미 문장을 구성하여 언어 자료로 만들었다. 남성 화자 1인이 이들 언어 자료를 3회 연속 발음하도록 하고 녹음하여 음성 자료로 채록하였다. 단기 자동상관기법을 사용하여 10차 선형 예측 계수와 운율 정보를 구해 도시하고, 이를 근거로 음소 분할을 실시하였다. 분할된 각 음소의 운율 변화 곡선을 다항식으로 근사하기 위해 비선형 곡선 정합 방법을 적용하여 초기 치와 다항식 계수를 구해 인공 신경망을 학습시키기 위한 운율 자료를 구축하였다.

인공 신경망의 학습 단계에서는 3회 발성된 자료 중 처음 2개의 자료를 학습에 이용하였는데, 입력 층에는 문장의 음소 열을 인가하고, 음소 열의 중앙에 해당하는 음소의 운율 정보를 출력 층에 목표 패턴으로 인가하여 인공신경망을 학습시켰다. 훈련 주기는 200회로 제한하고 그 전에 훈련을 마칠 수 있는 최소 오차 임계치를 설정하였다. 해당 음소에 대한 훈련이 끝나면 음소열을 왼쪽으로 이동시켜 다음 음소가 중앙 음소가 되게 하여 주변 음소 환경에 대한 초분절적 요인을 인공 신경망이 학습할 수 있도록 하였다. 언어 자료의 모든 문장에 대해 훈련을 마쳤을 때 각 음소에 대한 각 인공 신경

망의 추정율은 90 - 92%로 나타났다.

평가 단계에서는 입력 단에 문장의 음소 열을 인가했을 때 나타나는 인공 신경망의 출력 단의 값을 3번째 자료의 해당 음소의 피치 및 에너지에 대한 계수와 비교하여 구한 추정율은 89 - 90%이었다.

V. 결론 및 검토

피치 및 에너지 인공 신경망의 추정율이 학습 단계에서는 90 - 92%이고 평가 단계에서는 89 - 90% 였다.

인공 신경망의 추정율을 높이기 위해서는 먼저 언어 자료의 규모를 좀더 광범위하게 구축해야 할 것이다. 또한 현재 실험에서는 입력 단의 음소 수를 11개로 제한하고 있어, 중앙 음소의 전후 5개 음소의 영향은 제대로 반영할 수 있으나 그 이상의 영향을 제대로 반영할 수 없다는 문제점이 있다. 이러한 문제를 해결하기 위해서는 입력과 출력 노드 수를 늘리면 가능하겠으나 계산량이 기하급수적으로 늘어나는 문제가 있다.

언어 자료의 규모가 작으면 과도 학습의 문제도 발생할 수 있을 것이다.

참고문헌

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

[2] A. Waibel, *Prosody and Speech Recognition*. Morgan Kaufmann Publishers, 1938.

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

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

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

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

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

[8] C. Tuerk and T. Robinson, "Speech Synthesis Using ANN Trained on Cepstral Coefficients," in Proc. EUROSPEECH '93, 1993, pp.1713-1716

[9] M. Riedi, "A Neural-Network-Based Model of Segmental Duration for Speech Synthesis," in Proc. EUROSPEECH '95, 1996, vol.1, pp.599-602.

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

[11] 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, Vol. 1 of 2, pp. 207 - 210

[12] 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.

[13] Il-Goo Lee, Chan-Goo Kang, Joon-Sik Kim. Un-Cheon Lim, "Prosody Generator for Speech Synthesizer Using Artificial Neural Networks," in Proc. ICSP'99, 1999, Vol. 1 of 2, pp. 183 - 186

[14] 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.

[15] 신동엽, 민경중, 강찬구, 임운천, " 인공신경망의 운율발생에 관한 연구," 2000년도 한국음향학회 학술발표대회 논문집 제 19권 제1(s)호, pp. 87-90.

[16] 신동엽, 민경중, 강찬구, 임운천, " 문장단위 운율발생용 인공신경망에 관한 연구," 2000년도 한국음향학회 학술발표대회 논문집 제 19권 제2(s)호, pp. 53-56.

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

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

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

[20] 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.

[21] 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.