문장단위 운율발생용 인공신경망에 관한 연구

A Study on the Artificial Neural Networks for the Sentence-level Prosody Generation

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

요약

무제한 어휘 음성합성 시스템의 문-음성 합성기는 합성음의 자연감을 높이기 위해 여러 가지 방법을 사용하게되는데 그중 하나가 자연음에 내재하는 운율 법칙을 정확히 구현하는 것이다. 합성에 필요한 운율법칙은 언어학적 정보를 이용해 구현하거나, 자연음을 분석해 구한 운율 정보로부터 운율 법칙을 추출하여 합성에 이용하고 있다.

이와 같이 구한 운율 법칙이 자연음에 존재하는 운율 법칙을 전부 반영하지 못했거나, 잘못 구현되는 경우에는 합성음의 자연성이 떨어지게 된다. 이런 점을 고려하여 우리는 자연음의 운율 정보를 이용해 인공 신경망을 훈련시켜, 문장단위 운율을 발생시킬 수 있는 방식을 제안하였다.

운율의 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 단어 군과 음소 균형 문장 군을 화자로 하여금 발성하게 하여, 녹음하고, 분석하여 구한 운율 정보를 데이터베이스로 구축하였다. 문장 내의 각 음소에 대해 지속시간과 피치 변화 그리고 크기 변화를 구하고, 곡선 적응 방법을 이용하여 각 변화 곡선에 대한 다항식 계수와 초기치를 구해 운율 데이터베이스를 구축한다. 이 운율 데이터베이스의 일부를 인공 신경망을 훈련시키는데 이용하고, 나머지를 이용해 인공 신경망의 성능을 평가한 결과 운율 데이터베이스를 계속 확장하면 좀더 자연스러운 운율을 발생시킬 수 있음을 관찰하였다.

I. 서 론

인간이 가장 자연스럽게 통신할 수 있는 통신 수단중의 하나가 음성이다. 정보화 시대에 들어선 지금 컴퓨터가 인간의 말을 이해하고 처리하여 그 결과를 다시 음성으로 돌려줄 수 있다면 가장 편리한 정보 교환 수단이 될 수 있다. 이를 위해 1970년대부터 컴퓨터와 디지털 신호처리기술의 발달과 더불어 디지털 음성처리에 대한 연구가 본격적으로 시작되었다.

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 제안하였다. 문장 내의 각 음소의 피치와 크기 변화 곡선을 발생시키는 인공 신경망을 각각 설계하여, 학습시키고 그 성능을 평가하였다. 인공 신경망을 훈련

시키기 위해 고립 단어 군과 음소 균형 문장 군으로 구성된 언어 자료를 만들고, 이 언어 자료를 일정 환경에서 남성 화자 1인으로 하여금 3회 반복 발음하게 하여 녹음하고, 이것을 음성 자료를 만들었다. 작성된 음성 자료를 단기 분석하여 각 음소열에 대한 원시 운율 자료를 자성하였다.

원시 운율 자료내의 각 음소의 피치 변화와 크기 변화 곡선을 2차 다항식으로 근사하는 곡선 정합 방법에 의해 각 변화 곡선의 다항식 계수와 초기치를 구해 인공 신경망을 훈련시키고, 평가할 수 있는 운율 자료로 만들었다. 2장에서는 한국어의 운율에 대한 고찰과 언어자료 구축에 관해 논하였고, 3장에서는 운율 법칙을 훈련시키기 위한 인공 신경망에 대해, 4장에서는 실험 방법과 그 결과에 대해 기술하였다.

II. 한국어의 문장 단위 운율

운율의 3가지 요소는 각 분절의 지속시간, 피치 변화, 크기 변화로 이루어지며 이들 각 분절의 운율 정보는 각 분절 고유의 특징을 포함하기도 하나 다양한 주변 요인에 의해 변하게 된다. 특히 주변 분절에 의한 초분절적인 영향에 의해 각 분절의 운율은 변하게 된다. 이외에도 운율에 영향을 주는 것으로 화자의 개성이나 감정 상태 등 다양한 변화 요인이 있을 수 있다.

이 모든 변화 요인을 전부 반영하기 위해서는 광범위한 언어 자료와 발성 환경, 다양한 화자 등이 필요하게 되어, 막대한 시간과 노력이 필요하게 되므로, 본 논문에서는 화자의 개인적인 특징이 발성단계에서 개입되지 않도록 제한하여, 평정한 상태에서 문장을 발성하는 것으로 제한하였다. 구문론적인 측면에서는 실제 대화체 문장의 발음을 이용해 모델링할 수는 없기 때문에, 고립 단어와 평서문에서 구문의 구, 절 등의 경계와 단어의 강세 유형 그리고 분절에 의한 영향을 반영한 운율 법칙을 인공 신경망이 학습하게 하였다.

이러한 점을 감안해 본 연구에서는 음소 균형 고립단어 군과 문장 군을 신경망 훈련 및 평가를 위한 언어 자료로 구축하였다.

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

음성자료를 단기 분석하여 각 프레임별 10차 선형예측계수와 피치, 에너지를 구했고, 각 음소별로 분할하여 각 음소별 총 프레임 수, 피치 변화, 에너지 변화를 구해 운율에 대한 원시 자료로 만들었다.

각 음소의 지속시간과 피치 변화, 에너지 변화를 2차 다항식으로 근사하기 위해 곡선 정합 방법을 적용하여 초기치와 다항식 계수를 구해 신경망 훈련과 평가를 위한 운율 자료로 구축하였다.

III. 피치와 에너지 곡선 발생용 인공 신경망

음성 합성기의 합성 음질은 이해도와 자연감으로 평가하게 되는데, 연결 합성 방식에서는 합성단위를 다양하게 하여 운율 법칙의 정확한 구현이 어려운 분절간 천이부분 전부를 하나의 합성 단위로 사용하여 자연감은 크게 늘어났으나, 합성용 데이터베이스의 규모가 커지게되는 단점이 있다.

이와 같이 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 운율자료를 계속 늘려 가면 모든 가능한 경우의 운율 법칙을 학습시킬 수 있을 것이다.

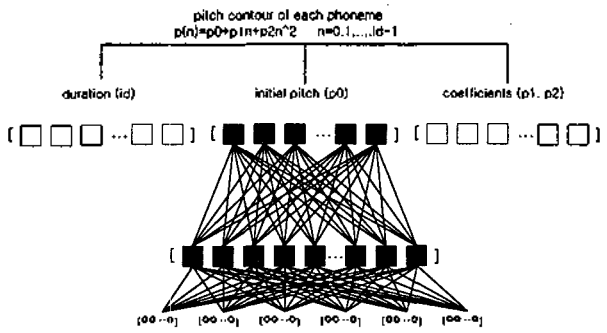

그림 1. 역전파 신경망의 구조
Fig.1 An Architecture of BP network

피치 변화와 에너지 변화를 학습하고 발생시키는 인공 신경망으로 역전파 신경망을 사용하였다. 그림 1. 은 피치 변화를 학습하여 발생시키는 인공 역전파 신경망의 구조이다. 에너지 변화 신경망도 동일한 구조를 갖는다.

인공 신경망의 입력으로 문장의 음소 열이 주어

진다. 은닉층은 한 층을 사용하고, 출력 층은 피치 발생 신경망의 경우 해당 음소의 피치 변화 곡선의 다항식 계수와 초기치, 지속 시간을 출력한다.

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. 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 입력 단에 포함된다면 인공 신경망을 더 효율적으로 학습시킬 수 있을 것이다.

한 운율 구 내의 음소 분절의 수가 2개에서 10개 이상까지 변하므로 초분절적인 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 은닉층을 사용하였고 은닉 층의 노드의 수는 입력 층의 노드 수와 같게 지정하였다.

출력 층은 입력 층의 중앙에 해당하는 음소에 대한 2차 다항식 계수를 출력한다. 다항식 계수와 초기치, 지속시간을 출력하므로 4개의 모듈로 구성된다. 각 다항식 계수와 초기치, 지속시간에 각각 16비트씩을 할당하였다.

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 선형예측계수 변화곡선을 이용하여 각 음소로 분할한 결과, 각 음소의 지속시간이 1 프레임에서 24 프레임까지 변화하는 것으로 나타났다. 각 음소의 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.

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

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

여기서 p_1, p_2는 피치 계수, p_0는 피치 초기치, d는 지속시간(프레임 수)이다. e_1, e_2는 에너지 계수이고, e_0는 에너지 초기치이다.

그림 2는 음소 '예'의 피치변화 곡선과 곡선정합 방법에 의해 구한 다항식 근사 곡선을 표시한 것이다. 그림 3은 동일 음소의 에너지 변화곡선과 그 추세선을 표시한 예이다.

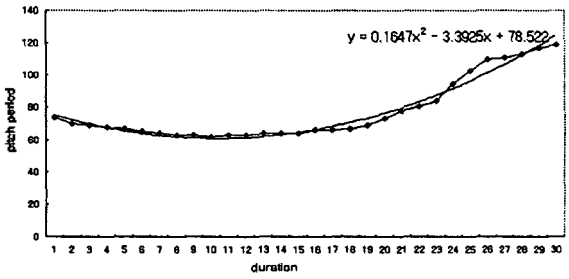

Fig. 2 Pitch contour of '예' phoneme and approximant line

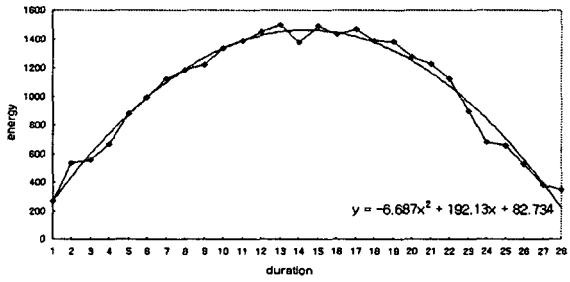

Fig. 3 Energy contour of '예' phoneme and approximant line

IV. 실험

인공 신경망을 훈련시키고 평가를 하기 위해 음소 균형 412개의 고립단어를 기반으로 100개의 의미 문장을 구성하여 언어자료로 만들었다. 남성화자 1인이 이들 언어자료를 3회 연속 발음하도록 하고 녹음하여 음성 자료를 채록하였다. 단기 분석 기법을 사용하여 10차 선형 예측계수와 운율 정보를 구해 도시하고, 이를 근거로 각 음소를 분할하였다. 분할된 각 음소의 운율 변화 곡선을 2차 다항식으로 근사시키기 위해 비선형 곡선 정합 방법을 적용하여 초기치와 다항식 계수를 구해 인공 신경망을 학습시키기 위한 운율 자료를 구축하였다.

인공 신경망의 훈련 단계에서는 3회 발성된 자료 중 처음 2개의 자료를 이용하였는데, 입력 층에 문장의 음소 열을 인가하고, 음소 열의 중앙에 해당하는 음소의 운율 정보를 출력 층에 목표 패턴으로 인가하여 신경망을 학습시켰다. 훈련 주기는 200회로 제한하고 그 전에 훈련을 마칠 수 있는 최소 오차 임계치를 설정하였다.

평가 단계에서는 입력 단에 문장의 음소 열을 인가했을 때 나타나는 인공 신경망의 출력 단의 값을 3번째 자료의 해당 음소의 피치 및 에너지에 대한 다항식 계수와 비교하여 추정율을 계산하였다. 운율 변화에 대한 인공 신경망의 학습 효과는 상당히 높게 나타났다.

V. 결론 및 검토

각 신경망의 추정율이 훈련 단계에서는 90 - 92%이고 평가 단계에서는 89 - 90% 였다. 추정율을 높이기 위해서는 우선 언어자료를 좀더 광범위하게 구축해야 하고, 입력 단의 음소 수가 11개로 제한되어 있는데, 한 운율 구의 음소의 수가 이 것을 벗어나면 그 영향을 제대로 반영할 수 없다는 문제점이 있다.

근사 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. 언어 자료가 부족하면 과도 학습의 문제도 발생할 수 있다.

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, 1988.

[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.I, pp.599-602.

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

[15] 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

[16] 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.