인공 신경망의 한국어 운율 발생에 관한 연구

민경중, 임운천
호서대학교 대학원 전자공학과

The Study on Korean Prosody Generation using Artificial Neural Networks.

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

요약

한국어 문-음성 합성 시스템(TTS: Text-To-Speech)은 합성음의 자연스러움을 증가시키기 위해 운율 발생 알고리듬을 만들어 시스템에 적용하고 있다. 운율 법칙은 각국의 언어에 대한 언어학적 정보나 자연음에서 구한 운율에 대한 지식을 기반으로 음성 합성 시스템에 적용하고 있다. 그러나 이렇게 구한 운율 법칙이 자연음에 존재하는 모든 운율 법칙을 포함할 수도 없고, 또 추출한 운율 법칙이 틀린 법칙이라면, 합성음의 자연감이나 이해도는 떨어질 것이므로, TTS의 실용화에 장애가 될 수 있다.을

이러한 점을 감안하여 본 논문에서는 자연음에 내재하는 운율을 학습할 수 있는 인공 신경망을 이용한 운율 발생 신경망을 제안하였다. 훈련단계에서 인공 신경망의 입력 단에 한국어 문장의 음소 열을 차례로 이동시켜 인가하면 입력 단의 중앙에 해당하는 음소의 운율 정보가 출력되도록 훈련시킬 때, 목표 패턴을 이용한 감독학습을 통해, 자연음에 내재하는 운율을 학습하도록 하였다. 평가 단계에서 문장의 음소 열을 입력하고, 추정율을 측정하여 인공 신경망이 한국어 문장에 내재하는 운율을 학습하여 발생시킬 수 있음을 살펴보았다.

ABSTRACT

The exactly reproduced prosody of a TTS system is one of the key factors that affect the naturalness of synthesized speech.

In general, rules about prosody had been gathered either from linguistic knowledge or by analyzing the prosodic information from natural speech. But these could not be perfect and some of them could be incorrect.

So we proposed artificial neural network(ANN)s that can be trained to learn the prosody of natural speech and generate it. In learning phase, let ANNs learn the pitch and energy contour of center phoneme by applying a string of phonemes in a sentence to ANNs and comparing the output pattern with target pattern and making adjustment in weighting values to get the least mean square error between them. In test phase, the estimation rates were computed. We saw that ANNs could generate the prosody of a sentence.

I. 서 론

인간과 기계 사이의 가장 효율적인 통신 수단의 하나인 음성을 사용하기 위해서는, 컴퓨터가 음성으로 된 명령을 인식하고, 그 명령을 수행한 다음 그 결과를 다시 합성음으로 알려주어야 한다. 이를 위해 컴퓨터 내에 음성 인식 및 합성 시스템을 설치하여 음성을 인식하고, 합성음을 발생시킬 수 있어야 한다.

대부분의 TTS 시스템은 합성음의 자연스러움을 증가시키기 위해 운율 발생 알고리듬을 만들어 TTS 시스템에서 사용하고 있다. 운율 법칙은 각 국의 언어에 대한 언어학적 정보나 자연음에서 구한 운율에 대한 지식을 기반으로 음성 합성 시스템에 적용하고 있다. 그러나 이렇게 구한 운율 법칙이 자연음에 존재하는 모든 운율 법칙

을 포함할 수도 없고, 또 추출한 운율 법칙이 틀린 법칙이라면, 합성음의 자연감이나 이해도는 떨어질 것이므로, TTS의 실용화에 장애가 될 수 있다. 또 새로운 운율 법칙이나 기존 법칙을 변경시키려고 할 때는 항상 알고리듬 전체를 수정해야하는 어려움도 따르게 된다.

이러한 문제를 해결하는 방안으로 문장 내의 운율 법칙을 학습할 수 있는 인공 신경망을 제안하여 한국어 문장 내의 각 음소의 운율을 학습하여 발생시킬 수 있는 인공 신경망의 구조에 대해 고찰하였다.

인공 신경망을 훈련시키기 위해 먼저 음소 균형 문장군으로 구성된 언어 자료를 구축하였고, 이 언어 자료를 일정 환경에서 남성 화자 1인으로 하여금 3회 반복 발성하게 하여 녹음하여, 이 녹음된 음성을 기본으로 음성 DB를 구축하였다. 구축된 음성 자료를 대상으로 단기 선형예측분석을 행하여 각 음소에 대한 원시 운율 자료를 구했다.

원시 운율 자료 내의 각 음소의 피치와 크기의 변화를 각각 피치와 에너지 곡선으로 표현할 수 있으므로, 이들 곡선을 다항식으로 근사하게 되면, 다항식 계수와 지속시간 만으로 음소의 운율 정보를 표시할 수 있다.

곡선 정합 방법을 이용해 각 변화 곡선의 다항식 계수와 초기 치를 구해 운율 패턴 DB를 구축하고, 이중 2개 집단은 피치 및 에너지 인공 신경망을 훈련시키는데 사용하고, 1개 집단을 이용해 인공 신경망의 성능을 평가하도록 하였다.

2장에서는 한국어의 운율에 대한 고찰과 언어자료 구축에 관해 논하였고, 3장에서는 운율 법칙을 학습하는 피치 및 에너지 인공 신경망에 대해, 4장에서는 실험 방법과 그 결과에 대해 서술하였다.

II. 한국어 문장 내에서의 운율

한국어 문장 내의 각 분절의 운율 정보는 각 분절 고유의 특징을 나타내면서도 다양한 주변 요인에 의해 변하게 된다. 특히 인접 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 요인에 따른 정보를 전부 추출하기 위해서는 광범위한 언어 자료와 발성 환경, 다양한 화자 등을 감안해야 하므로, 언어 자료 구축, 분석, 훈련 등에 막대한 시간과 노력이 필요하게 되어, 본 논문에서는 화자의 개인적인 특징이 발성 단계에서 개입되지 않도록 하기 위해 평정한 상태에서 문장을 발성하는 것으로 제한하기로 하고, 그 한계 내에서의 운율 변화에 대해 논의하였다.

구문론적인 측면에서는 실제 대화체 문장의 발음대신 평서문에서 구문의 구, 절 등의 경계와 단어의 강세 유형 그리고 분절에 의한 영향을 반영한 운율 법칙을 인공 신경망이 학습하도록 하기 위해 음성학적으로 균형 잡힌 고립단어로 구성된 언어 자료로부터 추출한 단어를 이용하여 의미 있는 일반 문장과 구를 작성하여 실험에 필요한 언어 자료를 구축하였다.

한국어의 경우 한 문장 내에 몇 개의 운율 구가 존재하는 것으로 연구 조사되었다. 그리고 실제 운율의 변화는 문장 전체에 걸친 변화가 아니고 문장내의 운율 구 혹은 발화 구에 따라 변화하는 것으로 판단하여 언어 자료에 문장과 구를 포함시켰다. 실제 사용한 언어 자료에는 운율 구 내의 음소 분절의 개수가 2개에서 10개까지 다양하게 나타난 문장이나 구가 포함되어 있다..

언어 자료 구축이 끝나면, 이를 기반으로 무향실에서 특정 남성 화자 1인이 언어 자료를 3회 반복하여 발음하게 하고, 이것을 녹음하여 10 KHz로 표본화하여 양자화 한 음성 DB로 만들었다. 음성 자료는 10차 선형예측계수와 피치, 에너지를 구하는 자동 상관 방식의 단기 분석을 통해 각 프레임별 분석 계수 열로 나타난다. 이때 각 음소의 피치와 에너지의 프레임에 따른 변화가 해당 음소의 피치 및 에너지 변화곡선이 된다. 각 프레임의 표본 수를 256 표본으로 하고 128 표본씩 이동시켜 단기 분석을 행하였다.

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 크게 늘어나지 않음을 알 수 있었다.

문장 내의 각 음소를 분할하여 구한 각 음소별 총 프레임 수(지속시간)와 피치 변화 그리고 에너지 변화가 운율에 대한 원시 자료로 이용된다. 즉 각 음소의 지속시간과 피치 변화, 에너지 변화를 2차 다항식으로 근사하면 모든 변화 곡선을 초기 치와 2차 다항식 계수, 지속시간 등 4개 계수로 나타낼 수 있고 이를 인공 신경망이 학습할 수 있게 된다.

각 음소의 피치와 에너지 변화 곡선을 근사하기 위한 2차 다항식은 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 초기 값, d는 음소의 지속시간(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 1.에 한국어 모음의 피치 변화 곡선과 그 곡선을 2차 다항식으로 근사하여 구한 근사선을 나타내었다. 이러한 변화 곡선에 대해 곡선 정합 방법을 적용하여 초기 값과 다항식 계수를 구해 인공 신경망의 훈련과 평가를 위한 운율 자료로 구축하였다.

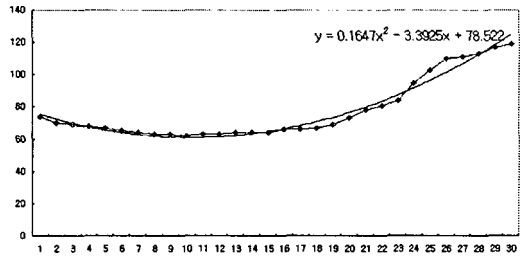

그림 1. 모음 /에/의 피치변화곡선과 그 근사선
Fig.1 The pitch contour of Korean vowel /ye/ and it's approximated line

III. 피치 및 에너지 인공 신경망

음성 합성 방식 중 무제한 어휘 음성 합성을 위한 법칙 합성 방식에서는 음소 단위의 운율 변화 법칙을 구현하여 합성음을 생성시키기 때문에 합성 단위 데이터베이스의 규모는 작은 대신 합성음의 자연감이 떨어진다. 합성음의 자연감과 이해도를 높이기 위해 합성 단위의 합성 계수에 대한 이해도 중요하나 운율이 자연스러운 변화를 하도록 하는 것도 중요하다.

이러한 운율 법칙은 언어학적 지식을 바탕으로 만들어질 수도 있고, 자연음에 내재하는 운율 법칙을 통계 분석을 통해 구할 수 있는데 두 가지 경우 다 모든 운율 법칙을 정확히 표현할 수 없다는 문제가 있다.

이런 점을 감안하여 인공 신경망으로 하여금 문장 내에 내재하고 있는 운율 법칙을 학습하도록 할 수 있다면 부정확하거나 구현하지 못한 운율 법칙을 인공 신경망에서는 정확한 운율법칙으로 학습하여 구현할 수 있을 것이다. 또한 언어 자료의 규모를 확대하고 발음 횟수를 늘려 훈련용 운율자료의 규모를 키우면 모든 가능한 경우의 운율 법칙을 학습시킬 수 있을 것이다.

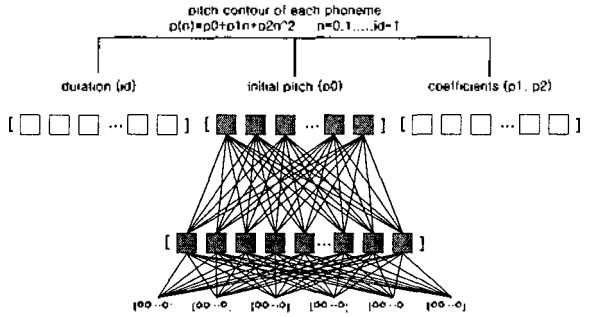

그림 2. 역전파 인공 신경망의 구조
Fig.2 Architecture of Back Propagation network

각 음소의 지속시간에 따른 피치 변화와 에너지 변화를 학습하여 발생시키는 인공 신경망으로 역전파 인공 신경망을 사용하였다. 그림 2는 피치 변화를 학습하여 발생시키는 피치 인공 신경망의 간단한 구조를 보여준다. 에너지 인공 신경망도 피치 인공 신경망과 동일한 구조를 갖도록 하였다.

각 인공 신경망은 입력 단에 문장의 음소 열을 인가하도록 설계되고, 은닉 층은 한 층을 사용하고, 출력 단에서는 피치 인공 신경망의 경우 해당 음소의 피치 변화 곡선의 다항식 계수와 초기 치, 지속 시간을 출력하고 에너지 인공 신경망의 경우 해당 음소의 에너지 변화 곡선의 다항식 계수와 초기 치, 지속 시간을 출력하도록 설계한다.

먼저 입력 단을 설계하기 위해 한국어 문장의 음운 변화를 마친 결과를 보면, 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 제시하고 인공 신경망은 주변의 전후 각 5 분절의 영향을 감안하여 학습하도록 설계하였다.

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

출력 층은 입력된 음소열 중 중앙에 해당하는 음소의 피치와 에너지 변화 곡선을 근사하는 2차 다항식 계수와 초기 치, 지속시간(프레임 수)을 출력하는 4개의 모듈로 구성하고 각 다항식 계수와 초기 치에 16비트 그리고 지속시간에 8비트를 할당하였다.

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 분석기법을 사용하여 10차 선

형 예측계수와 운율 정보를 구해 도시하고, 이를 근거로 각 음소를 분할하였다.

분할된 각 음소의 운율 변화 곡선을 다항식으로 근사시키기 위해 비 선형 곡선 정합 방법을 적용하여 초기치와 다항식 계수를 구해 피치 및 에너지 인공 신경망을 학습시키기 위한 운율 자료를 구축하였다.

인공 신경망의 훈련 단계에서는 3회 발성된 자료 중 처음 2개의 자료를 이용하였는데, 입력 단에는 문장의 음소 열을 인가하고, 음소 열의 중앙에 해당하는 음소의 운율 정보를 출력 층에 목표 패턴으로 인가하여 인공 신경망을 학습시켰다. 훈련 횟수는 200회로 제한하고 그 전에 훈련을 마칠 수 있는 최소 오차 임계치를 설정하였다. 각 인공 신경망은 91 - 92% 정도의 추정율을 보였다.

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 계산하여 90 - 91%의 성능을 나타내었다.

V. 결론 및 검토

각 인공 신경망의 추정율이 훈련 단계에서는 91 - 92%이고 평가 단계에서는 90 - 91% 이었다.

인공 신경망의 추정율을 높이기 위해서는 먼저 언어 자료의 규모를 좀더 광범위하게 구축해야 하고, 화자의 발성 횟수도 10회 이상으로 늘릴 필요가 있다. 언어 자료나 음성 자료의 규모가 작으면 과도 학습의 문제가 발생할 수 있다.

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 계산 부담이 기하급수적으로 늘어나는 문제가 있다.

인공 신경망에서 설정한 임계치와 훈련 횟수도 신경망의 성능에 영향을 줄 수 있을 것이다.

평가용 언어 자료와 다른 훈련용 언어 자료를 대상으로 인공 신경망의 추정율을 구해, 훈련 및 평가 자료가 동일한 문장을 기반으로 한 현재의 실험 결과와 비교하는 실험도 할 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, 1988.

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

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

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

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

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

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

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

[12] 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.

[13] 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.

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

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

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