음소별 성조 정보를 이용한 신경망 기반의 한국어 음소 지속시간 모델링

A Neural Network Based Korean Segmental Duration Modeling Using Tonal Information of Phonemes

김 은 경*, 이 상 호*, 오 영 환*
(Eun Kyoung Kim*, Sang Ho Lee*, Yung Hwan Oh*)

요 약

음소별 지속시간의 정확한 예측은 TTS 시스템의 자연성을 향상시키는데 중요한 역할을 한다. 기존의 한국어 음소 지속시간의 모델링을 위해 사용된 특징 변수에는 음소 문맥 정보, 품사 정보, 운율구 내에서의 위치 정보 등이 있다. 본 논문에서는 음소별 성조 정보 값을 새로운 특징 변수로 정의하여 예측 성능을 향상시키고자 한다. 성조 정보의 표현을 위해 두 개의 비경계 성조와 여섯 개의 경계 성조를 정의한 후, 400문장의 음성 코퍼스에 음절별 표기를 수행하였다. 성조 정보를 이용한 지속 시간 예측을 위해, 출력노드에서 음소의 지속 시간을 실수 형태로 출력하는 신경망을 구성하고, 이를 오류 역전파 알고리즘으로 학습시켰다. 실험 결과, 성조 정보를 사용하는 경우 실험 데이터에 대해 예측값과 실제값 사이의 상관계수로 0.863을 얻을 수 있었으며 이는 성조를 사용하지 않는 경우에 비해 향상된 성능을 나타내었다.

ABSTRACT

The accurate estimation of segmental duration is crucial for natural-sounding text-to-speech synthesis. For predicting Korean segmental durations, conventional methods utilized phonemic context, part-of-speech context and locational information in prosodic phrase. In this paper, the tonal information of phonemes is employed for more accurate prediction. After defining two non-boundary tones and six boundary tones, we annotated the tonal label on each syllable of 400 sentences. To predict segmental duration using tonal information, we constructed neural networks with a real-valued output node predicting phonemic duration and trained them by backpropagation algorithm. Experimental results showed that the proposed features are effective for predicting Korean segmental durations, and we got 0.863 correlation coefficient of the observed durations and predicted ones.

I. 서 론

문서 음성 변환(Text-to-Speech : TTS) 시스템은 임의의 문서를 입력으로 받아 그에 해당하는 음성을 출력하는 시스템으로, 인간이 자연스럽게 책을 읽는 것과 같은 수준의 음성을 출력하는 것을 목적으로 한다. 이러한 TTS 시스템의 평가기준은 크게 출력 음성의 명료성(intelligibility)과 자연성(naturalness)으로 나뉠 수 있으며, 현재 대부분의 TTS 시스템을 구성하는 단위음 연결 방식의 경우, 명료성은 우수한 반면 자연성이 떨어지는 경향이 있다. 따라서 최근에는 자연성을 향상시키기 위한 연구가 주를 이루는 실정이다.

일반적인 TTS 시스템은 크게 언어 처리부, 운율 생성부 그리고 음성 합성부로 나뉘어지며, 합성음의 자연성을 향상시키기 위해서는 언어 처리부에서 넘어오는 언어 정보로부터 정확한 운율이 생성되어야 한다. TTS 시스템이 제어하는 운율 요소에는 운율구 경계(끊어 읽기), 음소별 지속시간, 억양, 음의 세기가 있으며, 이들 각각이 정확히 모델링되어야 고품질의 자연성을 가지는 합성음 생성이 가능하다. 본 연구에서는 이 중 음소별 지속 시간의 예측 성능을 향상시키고자 한다. 음소 지속 시간은 억양과 함께 합성음의 자연성을 결정하는 가장 중요한 요소이며, 발성의 속도에도 영향을 받으므로 이를 제어하는 것이 자연성에 매우 큰 영향을 미칠 것이라 판단된다. 음소 지속 시간을 모델링하기 위한 기존의 방법으로는 CART(classification and regression trees)[1], sum of products model[2], 신경회로망[3] 등이 있다. 한국어의 경우, CART를 이용하는 트리 기반 모델링 기법과[4,5] 신경회로망을 이용한 방법을[6] 찾아볼 수 있다. 그러나 이러한 방법들에서 사용된 특징

* 한국과학기술원
접수일자 : 1999년 6월 17일

변수들을 살펴보면, 영어나 기타 다른 언어에서 사용되는 강세 정보가 배제된 것을 알 수 있다. 이는 한국어의 경우 강세의 표현 방법이 명확히 정의되지 않았기 때문이라 생각된다. 그러나 한국어의 경우에도 특유의 리듬이 존재하며[8], 영어나 기타 다른 언어의 강세 정보가 언어적 리듬감을 나타내는 요소라고 생각할 때, 한국어의 경우에도 성조 정보 즉 음성의 고저 정보를 이용할 수 있을 것이다.

본 연구에서는 한국어의 성조 정보를 특징 변수에 포함하여 신경망 기반의 음소 지속시간 예측 성능을 향상시키고자 한다. 2장에서는 사용된 코퍼스와 그의 처리를 설명하며, 3장에서는 본 논문에서 사용된 특징 변수들과 신경회로망의 구성을 설명한다. 4장에서는 성조 정보를 사용한 경우와 사용하지 않는 경우를 비교 실험하며 5장에서 결론 및 연구 방향을 제시한다.

II. 코퍼스의 구성 및 성조 정보의 표현

본 연구에서는 음소별 지속 시간 예측을 위해 초등학교 교과서, 논문 요약, 소설, 영화 해설 등에서 발췌된 단문, 복문, 평서문, 의문문, 감탄문 등의 400문장 (3,724어절)을 문장 코퍼스로 이용하였다. 또한 이를 전문 여성 아나운서가 방음실에서 발성한 약 33분 분량의 음성 코퍼스를 이용하였다.

문장 코퍼스는 기개발된 문서 분석기를 이용하여[7] 형태소 분석, 발음표기 변환, 구문 분석을 수행하고 분석 오류를 수정하였다. 또한 음성 코퍼스는 자동 음소 분할 프로그램을 이용하여 음소별로 분할하고 분할 오류를 수정한 후, 운율구 경계를 표시하였다. 표 1의 총 44개 음소를 이용하여 음성 코퍼스로부터 24,531개의 음소를 얻을 수 있었다.

표 1. 사용된 음소의 종류
Table 1. Class of phonemes.

조음방법	음소
단모음	ㅏ, ㅐ, ㅓ, ㅔ, ㅗ, ㅚ, ㅜ, ㅟ, ㅡ, ㅣ
이중모음	ㅑ, ㅕ, ㅖ, ㅘ, ㅙ, ㅛ, ㅝ, ㅠ, ㅢ
파찰음	ㅈ, ㅉ, ㅊ
마찰음	ㅅ, ㅆ, ㅎ
파열음	ㅂ, ㅃ, ㅍ, 종성 ㅂ ㄷ, ㄸ, ㅌ, 종성 ㄷ ㄱ, ㄲ, ㅋ, 종성 ㄱ
비음	ㄴ, ㅁ, 종성 ㅇ, 종성 ㅁ, 종성 ㄴ
유음	ㄹ, 종성 ㄹ

운율구 경계를 표시한 후, 음성 코퍼스의 성조(tone) 정보를 각 음절별로 표시하였다. 본 연구에서는 하나의 운율구내 억양이 연속된 비경계 성조(non-boundary tone)와 하나의 경계 성조(boundary tone)로 구성된다고 가정하고, 각 음절별로 하나의 성조 라벨을 갖게 하였다. K-ToBI[9]의 성조 표기법을 이용하여, 비경계 성조의 경우 H와 L, 즉 high 성조와 low 성조로 나타냈으며, 경계 성조는 L%, H%, LH%, HL%, LHL%, !HL%로 나타내었다. 경계 성조의 경우, K-ToBI에서 정의하고 있는 HLH%를 사용하지 않고 그 대신 !HL%를 사용하였는데, 이는 사용된 음성 코퍼스에서 HLH%가 발견되지 않았으며, LHL%의 경우에는 그 형태가 매우 다양하여 두 가지의 성조로 세분하는 것이 바람직하다고 판단되었기 때문이다. 그림 1의 피치 궤적과 같이, 같은 음절에 대해서 (a)와 같은 전형적인 LHL%의 궤적 이외에도 (b)와 같이 피치 궤적이 high 성조로 약간 올라가다가 급격히 low 성조로 내려가는 경우가 많이 발견되었으며 이러한 경우를 !HL%로 표기하였다. 한편, 모음과 자음의 각 성조별 평균 지속시간을 나타낸 그림 2에서 6개의 경계 성조 중 L로 끝나는 경계 성조의 경우가 H로 끝나는 경계 성조보다 지속시간이 긴 것을 알 수 있다. 이는 경계의 강약이 음소 지속시간 및 휴지기간의 길이에 의존하는 것으로 생각할 때, 운율적으로 약한 경계에서는 high 성조를 가지고 강한 경계에서는 high-low 성조를 가진다는 기존의 연구 결과와 일치한다[10].

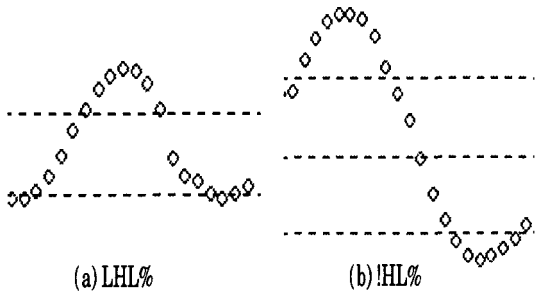

그림 1. 음절 '는'에 대한 LHL%와 !HL%의 예
Fig. 1. Examples of LHL% and !HL% for syllable 'nun'.

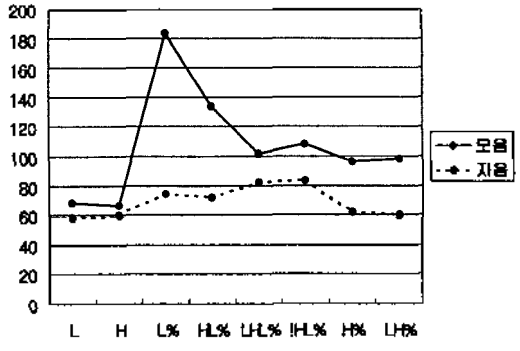

그림 2. 성조별 평균 지속 시간
Fig. 2. Average duration time of each tone label.

본 연구에서는 성능 평가를 위해 400문장 중 240문장 (15,037 음소)을 학습 데이터로, 160문장(9,494 음소)을 실험 데이터로 사용하였다. 각 데이터들은 여러 장르를 동일한 비율로 포함하도록 나뉘었고, 학습 데이터와 실험

데이터의 통계 특성이 전체 코퍼스의 통계 특성과 거의 일치하였다.

III. 신경회로망 기반의 한국어 음소 지속 시간 모델링

3.1 특징 변수

본 연구에서 음소 지속 시간 예측을 위해 선정한 특징 변수들은 다음과 같다. 각 특징 변수의 유효성을 알아보기 위해 카테고리 변수의 경우 F 값(F-ratio), 실변수의 경우 상관 계수(r)를 구하여, 표 2와 같은 결과를 얻었다. 성조 정보의 경우, 관측 성조값 뿐만 아니라 좌측 음소 및 우측 음소의 성조 값 모두 높은 F 값을 가짐을 알 수 있다.

● Dlph, Dcph, Drph : 관측 음소를 중심으로 좌측 음소, 관측 음소, 우측 음소를 나타낸다. 이 특징 값들은 음소의 문맥 정보를 나타내며, 관측 음소가 운율구의 처음 혹은 마지막에 위치하게 되면 각각 좌측 음소값과 우측 음소값이 NA (Not-Applicable) 값을 가지게 된다. 각 변수들의 F 값을 살펴보면, Drph의 경우 가장 높은 값을 나타냈는데, 이는 운율구 경계 앞에서의 장음화 현상에 기인한 것으로 생각된다.

● Dltone, Dctone, Drtone : 좌측 음소, 관측 음소, 우측 음소 각각의 성조값을 나타낸다. 음성 코퍼스의 성조 표기를 음절별로 수행하였으므로, 각각의 음절을 구성하는 음소들은 같은 성조 값을 갖는 것으로 가정하였다. 관측 성조를 나타내는 Dctone은 두개의 비경계 성조와 여섯 개의 경계 성조의 값을 갖게 되며, Dltone과 Drtone의 경우 NA값을 가질 수 있다.

● Dwhineoj, Dwhinphr : 어절 및 운율구 내에서의 음소 위치 정보로, 첫 음절, 중간 음절, 마지막 음절의 세 가지 값을 가질 수 있다.

● Cnsylineoj, Cnsylinphr : 어절내의 음절 개수와 운율구 내의 음절개수를 나타낸다. 일반적으로 운율구내의 음절 수가 많을수록 각 음절이 차지하는 시간이 짧아지며, 이러한 현상이 음소의 길이 예측에 유효한 값으로 작용할 것이라는 가정 하에 적용되었다.

표 2. 특징 변수의 F 값 및 상관 계수
Table 2. F ratio and correlation coefficient of features.

특징 변수	유효성 검사
Dlph	F=72.11
Dcph	F=227.40
Drph	F=351.18
Dltone	F=1255.89
Dctone	F=681.45
Drtone	F=1497.53
Dwhineoj	F=390.09
Dwhinphr	F=6283.59
Cnsylineoj	r=-0.02
Cnsylinphr	r=-0.06

한편 음소의 품사 정보는 기존의 한국어 음소 지속 시간 모델링에서 일반적으로 사용되는 특징 값이나, 품사 정보의 유효성을 알아본 실험 결과[4,6], 성능 향상에 도움을 주지 않는 특징 변수로 판명된 바, 본 연구에서는 품사 정보를 특징 변수로 사용하지 않았다.

3.2 신경회로망의 구성

본 연구에서는 하나의 은닉층과 하나의 출력 노드를 갖는 MLP(Multi-Layer Perceptron)를 이용하여 음소 지속 시간을 모델링하며, 학습 방법으로는 일반적인 오류 역전파(error backpropagation) 알고리즘을 사용하였다. 사용된 특징 변수 중 N개의 값을 갖는 카테고리 변수의 경우, N개의 노드들 중 하나만이 1의 값을 갖도록 입력 부호화 과정을 거쳤다. 3.1절의 특징 변수들 중에서 어절 및 운율구 내의 음절 개수를 제외한 나머지 특징 값은 카테고리 변수이므로 이들은 각각 이러한 부호화 과정을 거쳤다. 한편 MLP의 출력은 sigmoid 함수에 의해 0~1의 범위를 갖게 되므로, 출력값 또한 0~1의 값을 갖도록 부호화 과정을 거쳐야 한다. 본 연구에서는 sigmoid 함수의 특성을 고려하여 출력값을 0.01에서 0.99의 값을 갖도록 지속 시간을 선형 변환하여 실험하였다. 한편, 신경망의 과학습 문제를 해결하기 위해 학습 데이터 중 20%를 검증 데이터(validation data)로 사용하였다.

IV. 실험 및 결과

본 장에서는 성조 정보를 이용한 신경망 기반 음소 지속 시간 예측의 성능 평가와 비교를 수행한다. 신경망 기반의 예측을 위해 사용된 MLP는 3.2절의 입력 부호화 과정을 통해 성조 정보를 사용하지 않는 경우에는 146개의 입력 노드를, 성조 정보를 사용하는 경우는 173개의 입력 노드를 갖게 된다. 이를 은닉 노드의 수와 학습률을 변경해 가며 실험하여 가장 좋은 결과를 취하였다. 한편 전체 학습 데이터를 하나의 신경망으로 학습하는 경우, 모음과 자음으로 나누어 각각의 신경망을 학습하는 경우, 조음 방법에 따라 7개의 음소군으로 분류하여 학습하는 경우를 각각 실험하였다. 오류 평가 척도로 평균 제곱 오류근

표 3. 실험 데이터에 대한 결과
Table 3. Experimental results for test data.

학습방법	성조 정보 제외	성조 정보 포함
전체 학습 데이터	RMS : 20.57 RMSE : 0.285 CORR : 0.851	RMS : 19.70 RMSE : 0.261 CORR : 0.860
자음과 모음으로 분류	RMS : 20.30 RMSE : 0.277 CORR : 0.850	RMS : 19.62 RMSE : 0.259 CORR : 0.863
조음방법에 따라 분류	RMS : 20.48 RMSE : 0.282 CORR : 0.848	RMS : 19.74 RMSE : 0.263 CORR : 0.859

(root mean squared error : RMS), 상대 평균 제곱 오류 (relative mean squared error : RMSE), 상관 계수 (correlation coefficient : CORR)를 사용하여, 실험 데이터에 대해 표 3과 같은 결과를 얻었다.

표 3에서 보는 바와 같이 성조 정보를 사용함으로써 음소 지속 시간 예측의 성능 향상이 나타남을 알 수 있었다. 한편 전체 데이터를 하나의 신경망으로 학습하지 않고 예측 음소의 종류에 따라 나눠서 각각의 신경망으로 학습한 결과, 성능 차이가 거의 나지 않음을 알 수 있다. 이는 분류된 음소군 각각의 신경망을 학습하기에는 학습 데이터의 양이 부족했기 때문이라 판단된다.

그림 3은 표 3에서 가장 좋은 성능을 나타낸 성조 정보를 포함한 특징 변수 집합을 사용하고, 모음과 자음으로 분류하여 학습한 경우의 예측 성능을, 조음 방법에 따라 분류하여 나타낸 그림이다. 그림에서 보듯이, 모음의 경우가 자음의 경우에 비해 우수한 정확률을 보이며, 파찰음의 경우 다른 음소에 비해 매우 나쁜 성능을 보인다. 그림 4는 실제값과 신경망에 의해 예측된 값과의 차이를 V표식 상자-수염 그림으로 나타낸 것이다. 그림에서 보여지듯이 각 음소 클래스 오류들의 중앙값이 0에 가까움을 알 수 있다.

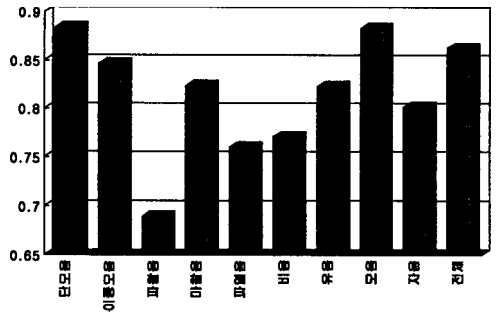

그림 3. 분류 음소별 예측 성능 (상관 계수)
Fig. 3. Prediction performance for each phonetic class (correlation coefficient).

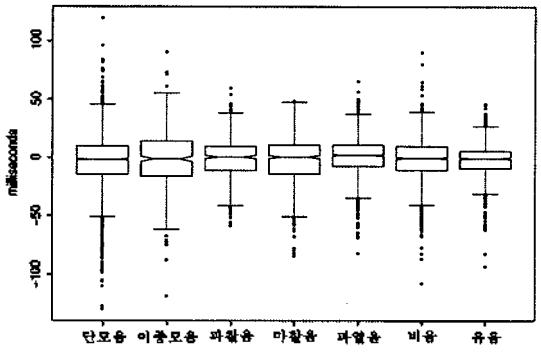

그림 4. 예측 오차값의 V표식 상자-수염 그림
Fig. 4. Notched box-and-whisker plot of residuals.

V. 결 론

본 논문에서는 한국어 문서 음성 변환 시스템을 위한 음소별 지속 시간을 예측하기 위해 음소별 성조 정보를 사용하였다. 성조 정보를 사용함으로써 예측값과 실제값 사이의 상관계수를 0.863까지 올릴 수 있었다. 조음 방법에 의해 분류된 음소를 각각의 신경망으로 학습한 경우, 하나의 신경망을 이용하는 경우에 비해 큰 성능 차이를 나타내지 않았는데, 이는 음소를 분류하는 과정에서 학습 데이터의 수가 줄어들었기 때문이라 판단된다. 추후 연구 과제로 본 논문에서 구현된 신경망을 실제 TTS 시스템의 음소 지속 시간 모듈로 이용하여 청각 테스트를 수행해야 할 것이다. 그리고 한국어에 적합한 특징 변수를 더 선정하는 과정이 필요할 것이다.

참 고 문 헌

1. M. D Riley, "Tree-based modelling of segmental duration," *Talking Machines: Theories, Models, Designs*, G. Bailly, C. Benoit and T.R. Sawallis, editors, pp. 265-273, Elsevier Science, 1992.
2. Jan P. H. van Santen, "Assignment of Segmental Duration in Text-to-Speech Synthesis," Computer Speech and Language, vol. 8 pp. 95-128, 1994.
3. Marcel Pazi Riedi, "Controlling Segmental Duration in Speech Synthesis Systems," PhD Thesis, Swiss Federal Institute of Technology, Zurich, 1998.
4. 이상호, 오영환, "운율구 추출 및 음소 지속 시간의 트리 기반 모델링", 한국음향학회지, Vol. 17, No 6, pp. 43-53, 1998.
5. 정지혜, 김인영, 이양희, "정규화 지속시간 회귀트리를 기반으로 한 음운지속시간 모델화", 한국음향학회 학술발표대회 논문집 제17권 2호, pp. 278-281, 1998.
6. 김온경, 이상호, 오영환, "한국어 문서 음성 변환 시스템을 위한 신경회로망 기반의 음소 지속시간 모델링", 한국정보과학회 봄 학술발표논문집(B) 제26권 1호, pp. 292-294, 1999.
7. 이상호, 오영환, 서정연, "한국어 문서 음성 변환 시스템을 위한 문서 분석기", 한국음향학회지, Vol. 15, No .3, pp. 50-59, 1996
8. 이현복, 한국어의 표준발음, 교육과학사, 1989.
9. M. Beckman and S.A. Jun, K-ToBI (Korean ToBI) Labeling Conventions, Ohio State University, 1996.
10. 김선미, 성굉모, "연속된 발화에서 운율구 내 high tone의 위치", 한국음향학회 학술발표대회 논문집 제 16권 2호, pp. 123-126, 1997.

▲김 온 경(Eun Kyoung Kim)
1995년 2월 : 한국과학기술원 전산학과(학사)
1997년 2월 : 한국과학기술원 전산학과(석사)

1997년 3월～현재 : 한국과학기술원 전산학과 박사과정 재학중
※주관심분야 : 음성합성, 운율제어, 패턴인식

▲이 상 호(Sang Ho Lee)
1993년 2월 : 동국대학교 전자계산학과(학사)
1995년 2월 : 한국과학기술원 전산학과(석사)
1995년 3월～현재 : 한국과학기술원 전산학과 박사과정 재학중
※주관심분야 : 음성합성, 자연언어처리, 패턴인식

▲오 영 환(Yung Hwan Oh)
1972년 : 서울대학교 공과대학(학사)
1974년 : 서울대학교 교육대학원(석사)
1980년 : Tokyo Instute of Technology 정보공학전공(박사)
1981년～1985년 : 충북대학교 컴퓨터공학과 조교수
1983년～1984년 : University of California, Davis 연구교수
1995년～1996년 : Carnegie-Mellon University 연구교수
1985년～현재 : 한국과학기술원 전산학과 교수
※주관심분야 : 음성인식, 음성합성, 음성코딩, 화자인식, 대화관리, 신경회로망, 전문가 시스템