한국어 대화체 TTS 개발을 위한 발음 및 운율 추정

이진식*, 김승원*, 김병창**, 이근배*
포항공과대학교 컴퓨터공학과*
대구가톨릭대학교 컴퓨터정보통신공학부**

Grapheme-to-Phoneme Conversion and Prosody Modeling for Korean Conversational Style TTS

Jinsik Lee*, Seungwon Kim*, Byeongchang Kim**, Gary Geunbae Lee*
Department of Computer Science and Engineering, POSTECH*
Computer and Information Communications Engineering, Catholic University of Daegu**
E-mail : palcery@postech.ac.kr, rockzja@postech.ac.kr, bckim@cu.ac.kr, gblee@postech.ac.kr

Abstract

In this paper, we introduce a method for extracting grapheme-to-phoneme conversion rules from the transcription of speech synthesis database and a prosody modeling method using the light version of ToBI for a Korean conversational style TTS. We focused on representing the characteristics of the conversational speech style and the experimental results show that our proposed methods are suitable for developing a Korean conversional style TTS.

I. 서론

기존의 한국어 TTS(Text-to-speech) 시스템 연구는 대량의 낭독체 말뭉치를 이용한 연결(concatenation) 기반의 방법론이 주를 이루었다 [1, 2]. 낭독체 말뭉치의 특징인 일정하고 안정된 목소리는 합성 DB와 동일한 도메인에 해당하는 입력에 대해서 합성 음질의 우수함을 보장해 줄 뿐만 아니라, 도메인을 벗어난 사람 이름이나 지명과 같은 고유명사의 합성에도 강건하다는 것이 특징이다. 하지만, 낭독체 합성 음성을 대화체 음성이 필요한 로봇 등의 안내 시스템에 적용하기에는 부자연스러운 측면이 있다. 따라서 보다 자연스러운 사용자와 시스템간의 의사소통을 위해서는 대화체 음성 합성 방법이 요구된다.

합성용 대화체 DB에서는 낭독체의 그것과는 달리 표준발음법과는 맞지 않는 발음이 종종 나타나기도 한다. 예컨대, "그럼요"의 경우, 표준발음법에 의하면 [그러묘]로 발음해야 되지만, 실제 합성 DB를 살펴보면, [그럼뇨]로 발음하는 경우가 있다. "어떻게"의 경우도 역시 [어떠케]로 발음하는 것이 옳지만, [어트케]로 발음하는 경우가 있다. 따라서 합성 음성이 합성 DB의 성격을 잘 따르도록 하기 위해서는, 합성 DB 음성 파일의 발음 전사(transcription)를 훈련 대상으로 하는 음소열-발음열 변환 모델이 필수적이다 [3].

대화체는 낭독체에 비해서 다양한 운율을 잘 표현할 수 있어야 한다. 이를 위해서는 충분히 많은 양의 음성 DB가 있거나, 제한된 음성 DB에서 만들어진 합성음의 높이, 세기, 길이 등에 수정을 가해야 한다. 하지만 충분히 많은 양의 음성 DB를 확보하는 것은 현실적으로 어려운 일이거니와, 합성 과정에서 기하급수적으로 늘어나는 계산량을 고려하지 않을 수 없다. 반면, 합성 음성에 수정을 가하게 되면 음질의 현저한 저하가 나타나는 단점이 존재한다. 물론, HMM 기반의 합성 방법이 그 대안이 될 수도 있지만, 이 방법론은 연결 기반의 방법론에 비해 현재까지는 합성 음성에 대한 평가가 좋지 못하다 [4]. 따라서, 제한된 음성 DB로 풍부한 운율을 표현할 수 있으면서도, 합성 음성의 수정을 최소화할 수 있는 방법론이 요구된다.

II. 시스템 전체 구성

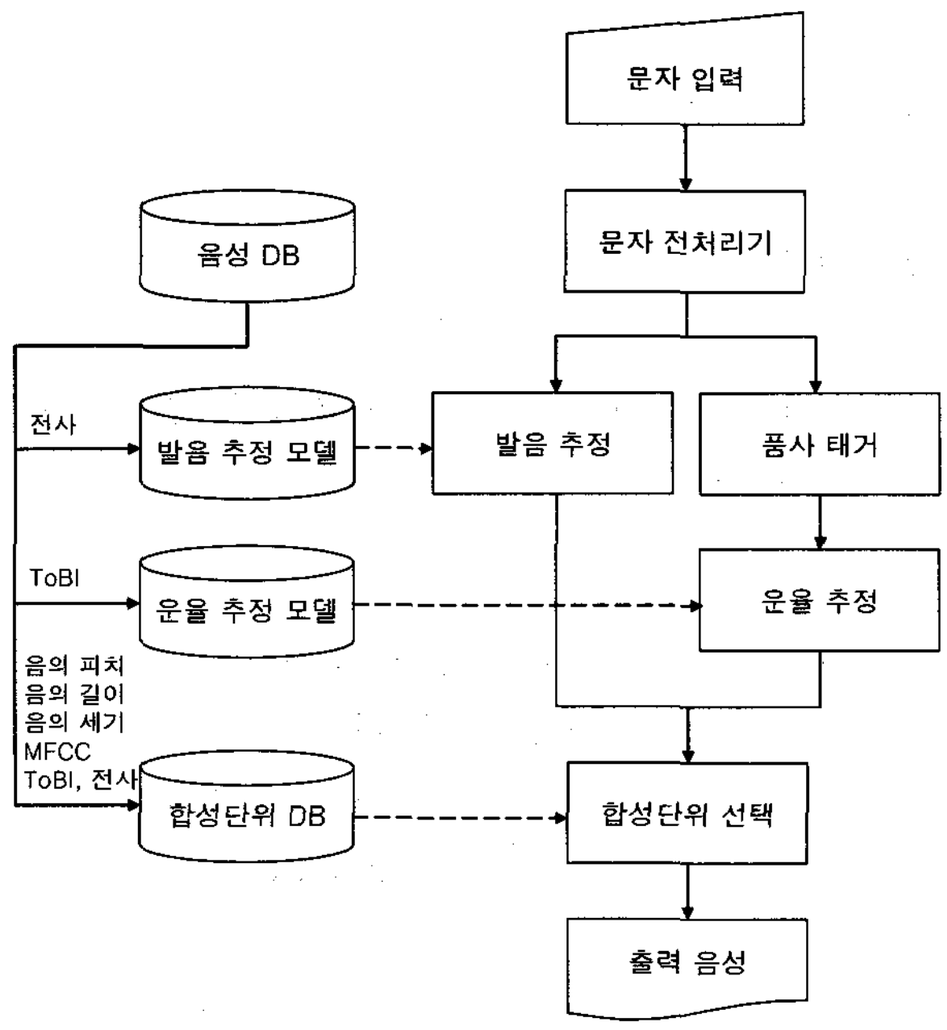

그림 1. TTS 시스템 전체 구성도

그림 1은 개발된 TTS 시스템의 전체 구성도를 나타내고 있다. 먼저 좌측의 발음 추정모델, 운율 추정모델, 합성단위 DB는 오프라인상에서 만들어진다. 먼저 발음 추정모델은 음성 DB의 발음 전사를 기반으로 훈련되며, 운율 추정모델은 수동으로 태깅된 ToBI로 훈련된다. 합성단위 DB는 음성 파일로부터 추출된 피치, 길이, 세기, MFCC 13차 계수와 수동으로 태깅된 ToBI 및 발음 전사 등의 정보를 담고 있다.

우측은 합성과정을 순서대로 도시한 것이다. 입력으로 들어온 문자열은 전처리기를 통해 시스템이 처리할 수 있는 형태의 내부 구조로 바뀐다. 앞서 만든 모델을 이용하여 발음과 운율을 추정하고, 추정된 결과를 바탕으로 가장 유사한 트라이폰(triphone) 합성단위를 선택하게 된다. 선택된 합성단위는 비용 함수(cost function)에 근거한 비터비 탐색(Viterbi search)을 통해 최적의 음성을 합성하는데 사용된다.

III. 발음 및 운율 추정 방법

3.1 음소열-발음열 변환

앞서 서론에서 언급되었듯이, 합성된 음성이 합성 DB의 특성을 따르기 위해서는 합성 DB의 발음 전사로부터 모델을 훈련시키는 것이 타당하다. 이 절에서는 발음 추정모델을 훈련하기 위한 방법과 훈련된 모델로부터 발음열을 생성하는 방법에 대해 기술한다.

(1) 음소열-발음열 정렬

한국어의 음절은 초성, 중성, 종성으로 이루어져 있다. 특히, 음소열의 종성이나 발음열의 초성, 종성의 경우 생략될 수 있기 때문에, 그림 2와 같이 각 음소를 세 자소(초성, 중성, 종성)로 쪼개어 음소열과 발음열을 정렬한다. "_" 문자는 생략된 자소를 나타내기 위한 기호이다.

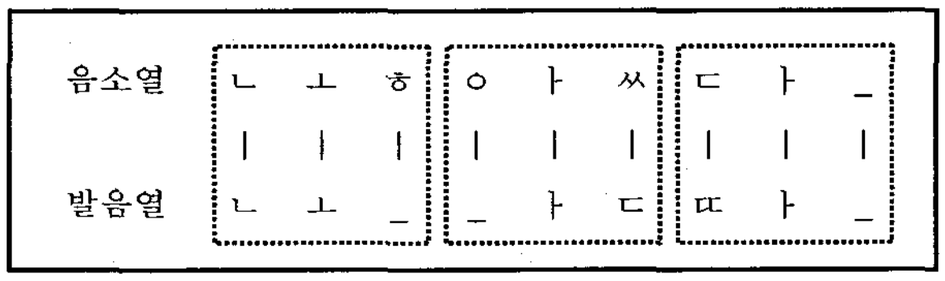

그림 2. 음소열-발음열 정렬의 예

(2) 규칙 생성

정렬된 음소열과 발음열을 이용하면 아래 (1)과 같은 규칙을 생성할 수 있다.

$$r\colon L(G)R \rightarrow P \qquad (1)$$

여기서 규칙 r은 왼쪽 문맥 L과 오른쪽 문맥 R을 만족하는 음소열 집합 G가 발음열 집합 P로 변환된다는 것을 의미한다. 이 때, L과 R의 길이는 가변적이며, G와 P는 자소 또는 "_" 문자로 이루어진 집합이다.

규칙 r은 하나 이상의 후보 발음열 $p \in P$를 가질 수 있는데, 이는 아래 (2)와 같이 실현 확률로 계산되어 그림 3의 규칙 트리에 저장된다. 그림 3의 "*" 문자와 "+" 문자는 각각 문장 경계와 어절 경계를 의미한다.

$$\Pr(p \mid L(G)R) = \frac{Count(L(G)R \rightarrow p \in P)}{Count(L(G)R \rightarrow P)} \qquad (2)$$

(3) 발음열 생성

발음열은 생성된 규칙 트리를 기반으로 후보 발음열 p 중에서 누적점수가 가장 높은 후보를 선택함으로써 생성된다. 누적 점수는 아래 (3)과 같이 계산된다.

$$Score(p \mid L(G)R) = \sum w_{CL} \Pr(p \mid L'(G)R') \qquad (3)$$

여기서, w_{CL}은 좌우 문맥 L'과 R'의 길이에 따른 가중치이며, L'과 R'은 각각 L과 R에 포함되는 문맥이다. 즉, 규칙 $L'(G)R' \rightarrow P$는 규칙 $L(G)R \rightarrow P$의 부모 규칙이거나 자기 자신에 해당한다.

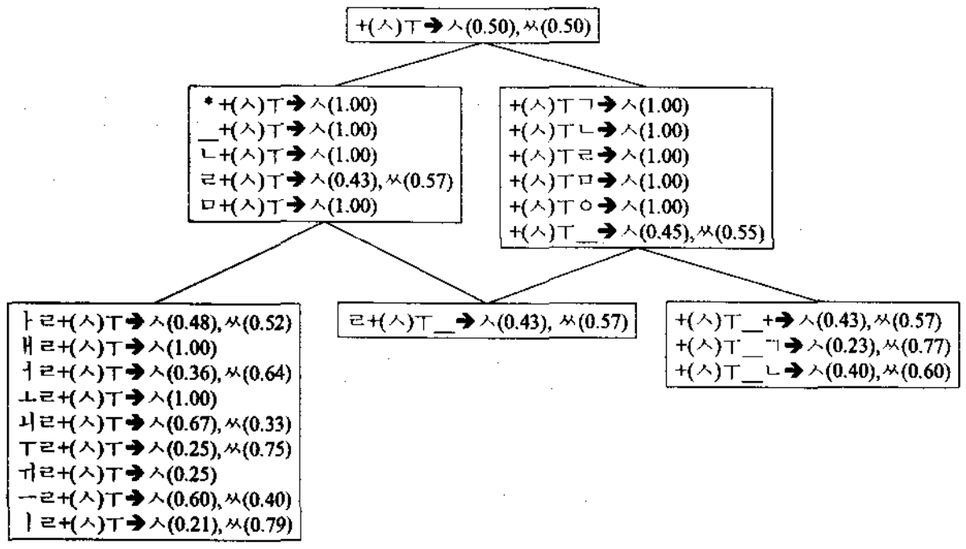

그림 3. 생성된 규칙 트리의 예

3.2 운율 추정

(1) 한국어 토비(Korean TOnes and Break Indices)

운율을 모델링하기 위해 운율 전사 규약인 한국어 토비를 사용하였다 [5]. 한국어 토비에는 다양한 톤과 경계 색인이 있지만, 본 논문에서는 이를 단순화하여 억양 구(Intonational Phrase)의 경계 톤 4가지(L%, H%, HL%, LH%), 악센트 구(Accentual Phrase)의 경계 톤 2가지(La, Ha)와 운율 경계 3가지(B_0-경계없음, B_2-소운율경계, B_3-대운율경계)만을 이용하였다.

(2) 운율 경계 추정

운율 경계는 문장의 운율 구조를 형성하기 때문에, 잘못 추정되었을 경우 본래 문장이 가지는 의미가 바뀔 수 있어, TTS 시스템에서 중요한 부분을 차지한다. 기존 연구에서는 운율 경계를 HMM(Hidden Markov Models) [6]이나 CART(Classification and Regression Trees) [7]를 이용해 추정했지만, 운율 경계를 추정하는데 중요한 요소 중 하나인 이전 운율 경계나 다음에 오는 운율 경계와의 거리는 고려하기 어려웠다. 따라서 본 연구에서는 운율 경계간의 거리를 고려하기 위해 ME(Maximum Entropy)를 기본 학습방법으로 하는 SSL(Stacked Sequential Learning) 기법을 추정에 사용하였다 [8]. 추정에 사용된 자질들은 표 1에 정의되어 있다.

(3) 톤 추정

낭독체 음성과 대화체 음성은 톤에서 가장 많은 차이를 보인다. 대화체에서는 같은 문장이라도 다양한 톤으로 발음될 수 있는데, 다양한 톤을 반영하기 위해 피치 곡선 전체를 추정하는 것은 어려운 일이다. 설사 피치 곡선을 잘 추정하였다고 하더라도 말뭉치 기반의 TTS 시스템에서는 추정된 피치에 해당하는 합성 단위가 부족하다는 한계가 있다.

본 연구에서는 대화체의 다양한 톤 변화가 운율 경계의 마지막 음절에서 주로 일어난다는 점에 착안하여 추정된 운율 경계의 마지막 음절에 대해서만 톤을 추정하였다. 톤의 추정은 CRF(Conditional Random Fields)[9]를 이용하였으며 사용된 자질은 표 2와 같다.

표 1. 운율 경계 추정에 사용된 자질들

자질	내 용
품사	앞, 뒤 형태소 품사 5개
음절 어휘	앞, 뒤 음절 어휘 1개
어절 길이	앞, 뒤 어절 길이 2개
문장내 위치	문장 시작 및 끝에서부터의 음절 개수
대운율경계 분포	대운율경계간의 음절 개수 분포를 반영한 가중치

표 2. 톤 추정에 사용된 자질들

자질	내 용
품사	앞, 뒤 형태소 품사 5개
음절 어휘	앞, 뒤 음절 어휘 1개
운율 경계	추정된 운율 경계
어절 길이	현재 어절의 음절의 개수
문장내 위치	문장 시작 및 끝에서부터의 음절 개수

Ⅳ. 실험 및 결과

4.1 합성 DB 정보

본 연구에서는 ETRI에서 2003년도에 배포한 음성합성용 대화체 DB를 훈련 말뭉치를 사용하였다. 이는 한영 회화 책에서 추출된 대화체 13,630문장, 64,601어절로 구성되어 있으며, 문장에 대한 발음 전사와 함께 어절 마지막 음절에 단순화된 K-ToBI를 수동으로 달았다.

4.2 합성 음질 평가

합성 음질의 평가를 위해 본 연구실이 가지고 있는 낭독체 TTS 시스템 [1]과 본 논문에서 제안한 방법으로 구현한 대화체 TTS 시스템을 이용하여 각각 낭독체 10문장과 대화체 10문장을 합성하였다. 합성된 문장은 8명의 평가자에 의해 MOS(Mean Opinion Score) 테스트와 CCR(Comparison Category Rating) 테스트 거쳤으며 그 결과는 표 3, 4에 요약하였다.

표 3. MOS 테스트 결과

테스트 문장	합성 시스템	Natural-ness	Clearness	Overall quality
대화체	대화체	4.2	4.5	4.4
	낭독체	3.0	3.1	3.2
낭독체	대화체	2.2	2.5	2.6
	낭독체	3.6	3.7	3.6

표 4. CCR 테스트 결과 (대화체 TTS의 선호도)

테스트 문장	매우 좋음	좋음	약간 좋음	비슷함	약간 나쁨	나쁨	매우 나쁨
대화체	5	1	1	0	1	0	0
낭독체	0	0	2	0	4	2	0

표 3의 MOS 테스트 결과에서는 본 연구에서 구현된 대화체 TTS 시스템이 대화체 문장에 대해서는 잘 동작하지만, 낭독체 문장에 대해서는 만족할만한 성능을 보여주지 못하고 있다. 반면, 기존의 낭독체 TTS 시스템은 대화체 문장과 낭독체 문장에 크게 상관없이 일정한 성능을 보여준다. 표 4의 CCR 테스트 결과는 기존의 낭독체 TTS 시스템과 비교하여 대화체 TTS 시스템의 상대적인 선호도를 평가한 결과이다. 이는 대화체 TTS 시스템이 낭독체 문장에 대해서는 다소 부적합하지만, 대화체 문장에 대해서는 매우 선호되는 경향을 보여준다.

V. 결론

본 논문에서는 대화체 문장에서 나타나는 발음 변이와 다양한 운율을 적절히 표현할 수 있는 대화체 TTS 시스템 개발을 위한 발음 및 운율 추정 방법을 소개하였다. 합성 DB의 발음 전사로 훈련시킨 음소열-발음열 변환 모델을 통해 합성 DB의 성격에 맞는 발음열을 생성하였으며, 한국어 토비를 이용한 운율 경계 및 톤 추정 모델은 말뭉치 기반 TTS 시스템의 부족한 합성 단위의 한계를 극복하면서도 자연스러운 대화체 운율을 생성해 낼 수 있었다. 하지만 현재의 대화체 TTS 시스템은 낭독체 문장을 대화체 느낌을 살려 합성하지는 못하고 있어서, 향후 연구는 대화체 음성 특징을 보다 잘 표현할 수 있는 한국어 대화체 TTS 시스템 개발을 목표로 하고 있다.

감사의 글

본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었다. (IITA-2005-C1090-0501-0018)

참고문헌

[1] 김병창, "한국어 TTS를 위한 발음 및 운율 생성," *박사학위논문*, 포항공과대학교, 2002.

[2] K. Yoon, "Building A Prosodically Sensitive Diphone Database For Korean Text-to-speech Synthesis System," *Ph.D. Thesis*, The Ohio State University, 2005

[3] J. Lee, S. Kim and G. G. Lee, "Grapheme-to-Phoneme Conversion Using Automatically Extracted Associative Rules for Korean TTS System," In *Proceedings of ICSLP*, pp. 1264-1267, 2006.

[4] S. Kim, J. Kim and M. Hahn, "Implementation and Evaluation of an HMM-Based Korean Speech Synthesis System," *IEICE Trans. on Information and Systems*, Vol. E89-D, No .3, pp. 1116-1119, 2006

[5] S. Jun, K-ToBI (Korean ToBI) labeling conventions (version 3.1), http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html, 2000

[6] P. Taylor and A. W. Black, "Assigning Phrase Breaks from Part-of-speech Sequences," *Computer Speech and Language*, Vol. 12, No. 4, pp.99-117, 1998.

[7] K. Yoon, "A Prosodic Phrasing Model for a Korean Text-to-speech Synthesis System," *Computer Speech and Language*, Vol. 20, No. 1, pp.69-79, 2006.

[8] S. Kim, J. Lee, B. Kim and G. G. Lee, "Incorporating Second-Order Information Into Two-Step Major Phrase Break Prediction for Korean," In *Proceedings of ICSLP*, pp.2370-2373, 2006.

[9] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," In *Proceedings of the ICML*, pp.591-598, 2000