한국어 합성음 평가 가이드라인 시안

조철우, 이상호, 김수진
창원대학교, (주)첫눈, 나사렛대학교

Evaluation Guidelines of Korean Synthetic Speech

Cheolwoo Jo, Sangho Lee, Sujin Kim
Changwon National University, First Snow Ltd, Korea Nazarene University

cwjo@changwon.ac.kr, sangholee@1nooncorp.com, sjkim@kornu.ac.kr

Abstract

This paper suggests some guidelines on evaluating Korean text-to-speech systems in various aspects. Guidelines are suggested in terms of text analysis, intelligibility test and naturalness test and also in terms of generalities and specialties.

I. 개 요

TTS 시스템을 평가하는 작업은 TTS 개발자와 사용자 모두에게 있어서 매우 중요한 과정임에도 불구하고 현재까지 한국어 TTS 평가에 대한 체계적인 연구가 별로 보고된 바 없다. 다른 언어의 음성합성기에 관해서는 이미 다양한 형태의 합성음 평가법 표준화를 위한 연구와 표준화기준이 제시된 바 있다.[1][2][3][4][5] 이에 TTS 개발 경험을 가진 공학자와 음성 언어 학자들의 회의를 통해 본 보고서를 작성하기에 이르렀다.[6] 우선, 한국어 TTS 평가 가이드라인을 작성하기 위해 고려한 사항들은 다음과 같다.

1.1 주관적 평가 방법 선택

한국어 TTS 시스템의 평가는 객관적 평가와 주관적 평가로 나뉜다. 전자의 경우는 주로 TTS 시스템 개발자들이 개발 과정에서 얻어지는 제품 성능을 이야기하고, 후자의 경우는 시스템의 최종 사용자가 판단하는 블랙박스 평가 결과이다. 전자의 평가 방법은 정확한 수치를 제시한다는 점에서 장점을 가지지만, 1) 시스템 개발자가 개발 과정에서 얻어지는 값을 솔직하게 제시해야 되고, 2) TTS 음성 제공자끼리의 변이가 존재하므로 서로 다른 두 시스템의 객관적 평가 값들은 서로 비교하기가 어려우며, 3) 객관적 수치가 인간의 청각 시스템과 과연 얼마나 일치하는지 그 정도를 보장할 수 없고 지금까지 개발된 방법중 수용가능한 적절한 방법이 부족하므로 본 보고서에서는 평가 방법으로 자연성을 제외하고는 주관적 평가 방법만을 고려한다.

1.2 현재의 연구 개발 방법을 고려

TTS 시스템 개발은 1960년대 말의 Klattalk부터 시작하여 현재까지도 계속 연구 개발되고 있는 분야이다. 40년 정도의 개발 기간 동안 시스템 개발 환경이 계속 개선되면서 시스템의 개발 방법도 역시 많은 변화를 이루었다. 초기 rule-based 시스템으로부터 현재의 대부분을 차지하고 있는 코퍼스 기반 시스템까지 서로 다른 시스템 성능과 평가법을 요구하고 있다. 아울러 시스템의 평가 방법도 현재 주로 연구되는 개발 방법을 고려해야 한다. 향후 기술 발전에 따라 조음합성에 기반한 규칙합성 등 합성방식이 변화할 경우 이에 따라 평가방법도 개선될 필요가 있다.

1.3 한국어의 대표성과 TTS 시스템의 특수성 고려

본 연구의 목표는 한국어 TTS 시스템을 평가하는 것이다. 일반적으로, 평가 대상에 대한 공정한 평가 방법은 평가 대상의 대표적인 면을 올바르게 평가하는 것이 합당할 것으로 사료된다. 즉 한국어의 아주 일반적인 면에 대한 공정한 평가가 무엇보다도 우선되어야 된다는 점에서 이견이 없다. 한편, 시스템 개발자들은 이러한 점을 어느 정도 고려하며 개발할 것임이 확실하므로, 시스템 개발 측면에서 취약한 점을 찾아내어 그 부분을 어떻게 해결했는지 조사할 필요가 있다. 그러므로, TTS 시스템의 각 평가 항목에 대해 대표성과 특수성

두 부분으로 나누어 평가할 필요가 있다.

1.4 문서 분석, 명료도, 자연성, 세 항목에 대해 평가

일반적으로 TTS의 평가는 크게 합성음의 언어 정보 전달 능력을 평가하는 명료도 (intelligibility) 테스트와 합성음이 얼마나 인간 음성의 운율과 비슷한지를 평가하는 자연성 (naturalness) 테스트로 나뉜다. 한편, 최근에 문서의 형태가 점점 다양해지고 있으므로, 이를 고려한 문서 분석 (text analysis) 능력도 함께 평가되어야 한다. 그러므로 TTS는 이러한 세가지 분야에서 평가하기로 한다.

II. 문서분석 평가

2.1 대표성 평가

TTS 시스템은 다양한 형태의 입력 문서에 대해, 사람이 그 문서를 읽었을 때의 형식과 동일한 형식의 합성음을 출력하기를 기대하게 된다. 특히, 그 문서 내에 있는 기호와 숫자의 처리는 시스템마다 다를 수 있으나, 일반적으로 수용할 수 있는 수준의 문서 분석 결과를 제시해 주어야 할 것이다. 아울러 다양한 형식의 문서 파일에 대해서도 처리할 수 있는 능력이 기대된다. 아래의 내용은 이를 중심으로 고려한 사항이다.

2.1.1 숫자의 처리

1) 서수/기수의 선택
2) 일반 숫자 읽기
3) 전화번호 읽기
4) 년도/날짜 읽기
5) 시간/대회성적 읽기

2.1.2 외국어의 처리

이는 한국어 문서에서 발생될 수 있는 영어, 한자, 일어, 불어, 독어와 같은 외국어를 한국어 발음/혹은 외국어 발음으로 읽을 수 있는지를 평가한다. 가능한 한 TTS 합성음 제공자의 목소리로 외국어를 합성하기를 기대하며, 그 경우 시스템의 완성도가 더 높다고 평가한다. 위 언어 중, 적어도 영어와 한자는 현 한국어 쓰임에 있어서 중요도가 높으므로, 반드시 합성이 가능하여야 한다.

2.1.3 발음 변환의 정확성 평가

한국어 발음 변환이 얼마나 정확한가를 평가

1) 형태소 분석 정확성에 대한 결과
2) 예외 발음 사전의 정확성 조사 (coverage 조사)

2.2 특수성 평가

문서 분석의 특수성 평가는 1) 시스템 개발 과정에서 발생할 수 있는 시스템 취약 부분에 대한 처리 평가와, 2) TTS 시스템이 가질 수 있는 사용자 편이성에 관한 점을 고려한다.

2.2.1 파일 종류의 처리

다양한 문서 파일을 읽을 수 있는가를 평가한다.

2.2.2 파일 포맷의 처리

다양한 문서 포맷을 읽을 수 있는가를 평가한다.

2.2.3 저빈도 알파벳열 발음 처리

알파벳으로 쓰여진 비영어권 문자를 어떻게 발음하는가?

III. 명료도 평가

TTS 시스템의 명료도는 합성음이 가지고 있는 "언어 정보"가 얼마나 정확한 지를 조사하는 것이다. 여기에는 TTS 개발자가 개발 과정에서 얻어지는 여러 객관적인 척도를 함께 제출할 수 있지만, 1장에서 설명하였듯이, 합성음 자체에 대한 평가를 시도하는 주관적 방법을 사용하기로 한다.

3.1. 대표성 평가

본 절에서는 한국어의 대표적인 발음을 고려한 시스템의 명료도를 평가한다. 그러므로 합성음 개발에서 일반적으로 고려되는 고빈도 음소열이 평가 시료로 사용된다. 본 절에서 설명하는 명료도 평가 시료는 부록으로 함께 제시된다.

3.1.1 음절 합성의 명료도

단음절의 합성음 들려주고 이를 평가한다. 특히, "ㄱ, ㄷ, ㅂ" 세 종류의 종성 폐쇄음에 대한 구분이 명확한지를 조사한다. 예를 들어, 다음의 테스트 표를 만들고, 이를 N명의 피실험자에게 나눠준 후, 세 개 중 한 개씩 정답 합성음을 들려주어 얼마나 많은 사람들이 올바르게 들었는지를 평가한다. 전체 맞은 개수로 음절 합성의 명료도를 평가한다.

3.1.2 단어 합성의 명료도

합성음의 명료도를 조사할 수 있도록, 대응되는 두 단어에 대한 테스트 표를 만들고, 앞의 경우와 동일한 방법으로 조사한다. 이 때, 초성, 중성, 모음에 대한 각각의 대응 단어들을 만들어야 한다.

3.1.3 문장 합성의 명료도

문장 합성의 명료도를 알아보기 위해서 의미적으로 잘못된 문장을 합성하여 이를 피실험자로 하여금 받아 적게 한다. 총 다섯 개의 어절로 이루어지게 하고 그 어절들의 수식 구조는 다음과 같이 한다.

부사 --> 관형어 --> 주어 목적어 --> 서술어

예를 들어 다음과 같은 문장을 생각할 수 있다.

"높게 자는 생산이 춤을 푼다."

위 문장에서 부사와 관형어는 의미적으로 불일치하여야 하고, 관형어과 주어 역시 의미적으로 성립할 수 없게 구성한다. 주어와 목적어도 서로 함께 나타나기 힘든 말이며, 서술어 부분에는 목적어와 높은 확률로 공기(collocation)되는 단어와 발음이 비슷하나 다른 단어를 놓는다. 이렇게 함으로써 청자로 하여금 오류를 유발하게 만든다. 이렇게 구성된 문장들을 들려주고, 청자가 틀린 어절의 개수를 조사하여 문장 합성의 명료도를 조사한다.

3.2 특수성 평가

본 절에서는 최근에 주로 사용되는 TTS 시스템 개발 방법을 고려하여 설명한다. 최근에 주로 개발되는 시스템들은 대부분 코퍼스 기반 시스템이고, 이러한 방법론의 취약점은 해당 음소 문맥이 음성 DB에 존재하지 않을 때, 음소 연결 부분이 부드럽지 않다는 점이다. 시스템 개발자들은 대부분, 가능한 모든 음소 문맥을 포함하도록 음성 DB를 구성하지만, 한국어의 특성상, 샘플링하기 어려운 음소 문맥이 다수 존재한다. 본 절은 이러한 부분의 처리 방법을 어떻게 평가하는가에 대해 논하고자 한다.

3.2.1 저빈도 한국어 단어의 발음 평가

한국어 단어 중 매우 발생 확률이 낮은 단어를 수집하고, 이를 합성한 후 피실험자로 하여금 받아 적게 한다. 음성 시료 단어는 유의미한 단어로 선택하되 한국인의 일상적인 언어 생활에서 발생하기 힘든 단어이어야 한다. 최소 20명, 20단어 이상의 집합을 구성하여 실험한다. 이 때 주의할 점은 단어의 표기와 발음이 서로 상이하면 청자가 발음 변환 과정도 유추하여야 하므로, 단어의 발음 변이는 없는 것을 선택한다.

3.2.2 저빈도 외국어 단어의 발음 평가

알파벳으로 쓰여진 비영어권 문자의 음소 연결이 부드러운가를 평가한다.

Ⅳ. 자연성 평가

TTS 시스템의 자연성은 청자에게 들려주고 얻어지는 주관적 평가 값과, 개발자가 제시하는 객관적 평가 값으로 나뉘어 평가될 수 있다. 여기서는 전자의 경우를 중심으로 설명하고, 후자의 경우에 대해서도 개발자가 첨부할 수 있도록 한다.

4.1 대표성 평가

피실험자에게 한국어의 대표적인 문장을 합성하여 들려주고, 아래 항목의 설문지에 1점부터 5점까지 점수를 주도록 한다. 여기서 대표적인 문장이란, 문장의 구성이 한국어의 대표적인 운율을 나타낼 수 있어야 한다 (영어라면 문의 5형식 예가 모두 제시되는 정도). 운율에는 총 음의 경계, 음의 지속 시간, 억양, 강세가 있으므로, 각 운율 요소들에 대한 평가가 이루어지도록 문장 집합을 구성한다. 예를 들어 다음과 같은 요소를 고려한 문장을 생각할 수 있다.

- 음의 경계
- 음의 지속 시간
- 억양, 강세

이상과 같이 각 운율 요소를 고려한 문장을 조심스럽게 선별하여 문장 집합을 구성한다. 한편, 음성 시료에 대한 질문은 다음과 같다.

(1) 합성음에 잡음이 있는 것처럼 들렸는가?

(2) 합성음이 사람 목소리처럼 들렸는가?

평가는 5 단계로 평가하고 (5: 매우 자연스러움, 4: 자연스러움, 3: 보통, 2: 불편함, 1: 매우 불편함), 총 20개 이상의 음성 시료와 20명 이상의 피실험자를 구성하여 실험한다. 좀 더 자세한 질문을 생각해볼 수 있었으나, 일반인을 대상으로 하는 실험이므로, 질문을 아주 평범하게 선택하였음을 밝힌다.

4.2. 특수성 평가

빈도가 낮은 운율 현상을 포함하는 문장 합성 실험

최근의 TTS 시스템은 운율 예측 모듈이 코퍼스에 기반하여 학습된다. 그러므로, 일반적인 한국어 문장 패턴이 아닌 문장에 대해서는 부자연스러운 운율을 예측할 가능성이 높다. 이에 대한 평가를 함께 포함시킨다.

4.3 자연성에 대한 객관적 평가

본 절에서는 자연성을 객관적으로 평가하는 방법에 대해 논한다. 합성음의 자연성 정도와 밀접한 운율(prosody)은 크게 음의 경계, 음의 지속 시간, 음의 높낮이, 음의 크기로 나뉘어진다. 이 네가지 운율 요소에서 음의 경계 추출 (prosodic phrasing) 과정은 이산

(discrete)적인 단위로 표현되고, 나머지 운율 요소들은 각각 millisecond, Hz, dB로 나타내는 연속(continuous)적인 단위로 표현된다. 그러므로, 음의 경계 추출 부분만 다른 평가 방법이 사용된다.

1) 음의 경계 추출 평가

(1) 오류율 = (삽입 오류 개수 + 삭제 오류 개수)/전체 어절 개수

(2) 혼잡 행렬 (confusion matrix)

(3) 예측 정확률

2) 음소 지속 시간 (segmental duration), 음의 높낮이 (pitch), 음의 크기 (energy)

위 세 가지 운율 요소들은 단위 음성 제공자의 운율에 아주 의존적인 요소이므로, 서로 다른 TTS 시스템의 운율 요소를 동일한 평가 방법으로 비교한다는 것은 매우 어렵다. 즉, A 화자가 B 화자보다 운율 요소의 내재적 변이가 많다면, 동일한 모델링 기법으로 두 화자를 모델링할 때, B 화자에 대한 객관적 평가 결과가 더 좋게 나온다. 그러므로 조금이라도 비교를 좀 더 객관적으로 하기 위해서는 한 개가 아닌, 복수개의 평가 척도를 제시하고 이를 모두 관찰하여 좀 더 진실에 가까운 비교가 가능하도록 노력한다.

위 세 운율 요소들 중 음소 지속 시간은 millisecond 단위를 이용하고 모든 음소에 대해 평가한다. 음의 높낮이는 휴지부 (pause)를 제외한 나머지 음성에서 10 msec 마다 Hz 단위로 비교한다. 음의 크기는 역시 휴지부 (pause)를 제외한 나머지 음성에서 10msec마다 dB 단위로 비교한다. 한편, 음의 크기는 음의 높낮이와 아주 밀접한 상관도를 보이고 있고, 또한 전체 운율에 대한 영향도가 상대적으로 낮으므로, 비교 대상에서 제외되어도 전체 운율을 평가하는 것에 크게 영향을 미치지 않을 것으로 사료된다.

(1) 실제값과 예측치 간의 상관계수 (correlation coefficient)

(2) 상대평균제곱근 (relative mean squared error)

위에서 상관계수는 가장 쉽게 생각할 수 있는 척도로, 1.0은 완벽하게 예측치가 실제값과 일치하는 경우를 의미한다. 이 경우, 전술한 바와 같이 자료의 분산도에 따라 이 값이 바뀌므로, 상대평균제곱근을 함께 제시하도록 한다. 상대평균제곱근은 아래 수식과 같다.

$$RE = \frac{R(d)}{R(\mu)} = \frac{\frac{1}{N}\sum_{n=1}^{N}(y_n - d_n)^2}{\frac{1}{N}\sum_{n=1}^{N}(y_n - \mu)^2} \quad (1)$$

위 식에서 분모는 전체 자료의 분산을 의미하고 분자는 예측치 d에 대한 분산이다. 즉, 자료의 분산을 얼마나 줄였는지를 나타내는 척도로 상관계수와 함께 제시하면, 시스템의 성능을 이해하는 데 도움을 줄 것으로 사료된다. (y는 실제 값, N은 전체 자료의 개수, μ는 평균값) 위 평가 자료도 최소 100 문장 (1,000 어절) 이상의 학습에 참여 안 된 실험 자료에 대한 결과를 보여주어야 한다.

V. 맺음말

본 논문에서는 한국어 음성합성시스템을 평가하기 위한 가이드라인을 문장분석, 명료도 측정, 자연성 측정 등의 세 부분으로 나누어 제시하였다.

각각은 일반적인 특성을 다루는 대표성 및 개별 시스템의 특수성에 대한 항목으로 나누어서 평가 관점을 제시하였다.

본 논문에서 제시된 항목들은 기존의 평가기준들을 참고하여 한국어 합성음 평가에 적합하다고 생각되는 내용을 첨삭하여 작성하였으나 현재 실제 시스템에 대한 적용이 되지 않은 상태이므로 향후 실제평가에 따라 필요한 내용이 추가 또는 보완되어야할 것이다.

이곳에서 제시된 평가 단어, 어휘, 문장 목록[6] 및 평가항목표가 개발자, 사용자들에게 공유되어 합성기술의 발전과 표준화에 기여할 수 있기를 바란다.

감사의 말씀

본 연구는 SiTEC의 용역에 의해 대한음성학회의 위탁연구로 수행되었습니다.

참고문헌

[1] Shuichi Itahashi, Overview of the East-Asian Activities on Speech Corpora and Assessment, Delhi, India, 2004.

[2] Nick Campbell, Speech Output Systems Assessment: Following the Jenolan Synthesis Evaluation Workshop.

[3] Hiroyuki Nishi, Shuichi Itahashi, JEIDA Guidelines for Speech Synthesizer Evaluation, EALREW'98, 1998.

[4] Jialu Zhang, Shiwei Dong, Guidelines to Assessment of Speech Synthesis Systems for Chinese, EALREW'98, 1998.

[5] Chap.12 Assessment of Synthesis Systems, Handbook of Standards and Resources for Spoken Language Systems, pp.481-563, 1997.

[6] 조철우, 이상호, 김수진, 한국어 합성음 평가 가이드라인, 연구보고서, 대한음성학회, 2005. (인쇄중)