다국어 TTS 시스템 기반 한국어 음성 합성을 위한 텍스트 처리 파이프라인에 대한 구조적 고찰

박보경, 이건희
에이치디씨랩스

{bogyung, Gunhee_Lee}@hdc-labs.com

A Structural Analysis of Text Processing Pipelines for Korean Speech Synthesis Based on Multilingual TTS System

Bogyung Park, Gunhee Lee
HDC LABS

요 약

다국어 음성 합성(Text-to-Speech, TTS) 시스템은 주로 영어 데이터를 중심으로 설계되어 왔으며, 한국어 음성 합성에 적용할 경우 단순한 모델 미세조정만으로는 안정적인 합성 품질을 확보하기 어렵다. 특히 한국어는 형태소 단위 변형, 조사 결합, 숫자, 기호, 단위 표현 등 텍스트 해석 단계에서의 언어적 의존성이 높기 때문에 영어 중심의 텍스트 처리만으로는 합성 불안정성 및 발음 오류가 발생할 수 있다. 본 논문에서는 다국어 TTS 시스템 기반 한국어 음성 합성 과정에서 나타나는 텍스트 처리상의 주요 구조적 한계를 분석하고, 실제 입력 환경을 고려한 한국어 특화 텍스트 처리 파이프라인 구성 요소를 체계적으로 고찰한다. 이를 통해 영어 중심 다국어 TTS 시스템의 한국어 적용 과정에서 텍스트 해석 단계가 합성 품질과 안정성에 미치는 영향을 체계적으로 정리하고, 향후 한국어 TTS 및 에이전트 기반 음성 인터페이스 설계 시 고려해야 할 언어 처리 관점을 제시한다.

ABSTRACT

Multilingual Text-to-Speech (TTS) systems have been primarily designed and trained on English-centric data, making it challenging to achieve stable synthesis quality for Korean speech synthesis through model fine-tuning alone. Korean exhibits high linguistic dependency at the text interpretation stage, including morpheme-level variations, particle combinations, and the processing of numbers, symbols, and unit expressions. Consequently, relying solely on English-centric text processing can lead to synthesis instability and pronunciation errors. This paper analyzes the major structural limitations in text processing that arise when applying multilingual TTS systems to Korean speech synthesis and systematically examines the components of a Korean-specific text pipeline designed for real-world input environments. Through this analysis, we systematically identify how the text interpretation stage affects synthesis quality and stability in the process of adapting English-centric multilingual TTS systems to Korean, and present linguistic processing perspectives that should be considered in future Korean TTS and agent-based voice interface design.

Ⅰ. 서 론

음성 합성 기술은 딥러닝 기반 모델 발전과 대규모 학습 데이터로 자연스러운 음성 생성이 가능해졌으며, 하나의 모델로 여러 언어를 처리할 수 있는 다국어 TTS 시스템[1]은 효율성과 확장성 측면에서 실용적인 장점이 있다. 그러나 이러한 다국어 TTS 시스템의 상당수는 영어 데이터를 중심으로 설계되어, 영어와 언어적 특성이 상이한 언어에 적용할 경우 여러 한계가 드러난다.

한국어는 형태소 단위 변형이 빈번하고 조사 및 어미 결합을 통해 문법적 의미가 표현되는 교착어로, 텍스트 해석 단계에서 고려해야 할 언어적 요소가 많다. 특히 숫자, 기호, 단위 표현은 문맥에 따라 읽기 형태가 달라지며, 숫자에 결합하는 단위나 순서 표현에 의해 기수·서수 해석과 발음이 결정된다. 영어 중심의 텍스트 처리 가정을 그대로 적용할 경우 발음 오류나 운율 붕괴와 같은 합성 품질 및 안정성 저하가 발생할 수 있다.

더욱이 최근 대화형 음성 에이전트 및 LLM 기반 음성 인터페이스의 확산으로 정제된 문장뿐 아니라 비정형 텍스트가 TTS 입력으로 전달되는 사례가 증가하고 있다. 실제 서비스 환경에서는 숫자와 기호가 혼합된 표현, 약어, 혼합 언어 표현 등이 빈번하게 등장하며, 이러한 입력은 텍스트 해석 단계의 처리 방식에 따라 안정성에 큰 영향을 미친다. 한편 신경망 기반 또는 LLM 기반 텍스트 정규화 기법을 활용하여 이런 문제를 완화하려는

시도도 이루어지고 있으나, 추론 지연, 모델 크기, 온디바이스 적용 가능성 등 시스템 수준의 제약으로 인해 실제 음성 합성에 직접 적용하기에는 한계가 존재한다. 이에 따라 실제 TTS 시스템에서는 규칙 기반 처리와 경량 신경망을 결합한 하이브리드 구조를 포함한 텍스트 처리 파이프라인 설계가 여전히 중요한 고려 사항으로 남아 있다.

이러한 배경 속에서 기존 다국어 TTS 연구는 주로 음향 모델 구조나 학습 방법에 초점을 맞추어 왔으며, 한국어로 적용하는 과정에서 텍스트 처리 단계가 갖는 구조적 한계와 설계에 대한 논의는 상대적으로 충분히 다루어지지 않았다.

본 논문에서는 다국어 TTS 시스템 기반 한국어 음성 합성 과정에서 텍스트 처리 파이프라인이 갖는 구조적 한계를 분석하고, 실제 입력 텍스트를 고려한 한국어 특화 텍스트 처리 파이프라인 구성 요소를 정리한다. 이를 통해 다국어 TTS 시스템의 한국어 적용 과정에서 텍스트 처리 단계가 합성 품질과 안정성에 미치는 영향을 고찰하고, 향후 한국어 TTS 및 에이전트 기반 음성 인터페이스 설계 시 고려해야 할 언어 처리 관점을 제시한다.

Ⅱ. 본론

본론에서는 다국어 TTS 시스템 텍스트 파이프라인을 입력 특성, 텍스트 해석 단위, 처리 전략 관점에서 구조적으로 분해하고, 한국어 음성 합성에 요구되는 텍스트 처리 구성 요소를 단계별로 정리한다. 이러한 구조적 분해를 통해 영어 중심 다국어 TTS 시스템에서 암묵적으로 단순화되어 있던 텍스트 해석 단계가 한국어 환경에서는 어떠한 추가적 역할을 수행해야 하는지를 명확히 한다.

텍스트 처리 파이프라인은 입력 텍스트의 형태와 복잡도에 따라 서로 다른 처리 요구를 갖는다. 실제 서비스 환경에서는 숫자, 기호, 약어, 혼합 언어 표현이 포함된 비정형 텍스트가 빈번하게 입력되며, 이는 단순한 문자 변환을 넘어 입력 유형 인식과 처리 범위 구분이 필요함을 의미한다. 따라서 텍스트 파이프라인의 초기 단계에서는 언어 및 스크립트 인식, 비정형 요소 탐지와 같은 입력 분류 기능이 중요한 역할을 수행한다.

텍스트 해석 단위의 관점에서 보면, 영어 중심 다국어 TTS 시스템의 텍스트 처리는 주로 단어 단위의 정규화와 발음 변환을 중심으로 구성되지만, 한국어의 경우 숫자, 단위, 조사, 어미 결합 등으로 인해 해석 단위가 형태소 수준까지 확장될 필요가 있다. 특히 숫자 표현은 결합하는 단위 명사에 따라 고유어 수사 또는 한자어 수사가 선택되며 발음이 달라진다. "3명", "3개"는 "세 명", "세 개"로 읽히는 반면, "3층", "3회"는 "삼 층", "삼 회"로 발음되며, "3시간"은 "세 시간"이지만 "3분"은 "삼 분"으로 읽힌다. 이러한 규칙은 단순한 숫자 치환이나 토큰 단위 처리만으로는 안정적으로 반영되기 어렵다.

한국어는 조사와 어미 결합을 통해 문법적 의미가 표현되는 교착어로, 발음 단위가 형태소 경계를 따라 재구성된다. "집이", "집을", "집은"과 같이 조사에 따라 발음과 운율적 실현이 달라지므로, 발음 변환 이전 단계에서 형태소 분석[2]을 수행하고 이를 기반으로 발음 단위를 구성하는 과정이 필요하다.

혼합 언어 및 영어 약어 표현은 실제 입력 환경에서 빈번하게 등장한다. "AI 기술", "GPU 서버"와 같이 영어 약어와 한국어가 결합된 표현은 약어를 문자 단위로 읽을지 발음 단위로 확장할지에 대한 판단이 요구된다. 이러한 발음 해석 방식은 입력 표현의 성격과 사용 맥락에 따라 달라지므로, 언어 및 스크립트 인식 이후 혼합 언어 표현에 대한 별도의 처리 규칙이 필요하다.

텍스트 처리 전략의 관점에서 보면, 형식이 명확한 숫자, 기호, 단위 표현은 규칙 기반 처리로 안정적으로 처리할 수 있는 반면, 문맥 의존성이 높은 표현이나 혼합 언어 입력은 보다 유연한 해석이 요구된다. 이에 따라 한국어 음성 합성을 위한 텍스트 처리 파이프라인은 규칙 기반 처리와 경량 신경망 기반 처리[3][4]를 결합한 하이브리드 구조로 구성될 수 있으며, 이는 처리 지연과 시스템 복잡도를 최소화하면서 다양한 입력 조건에 대응할 수 있는 현실적인 설계 방향을 제시한다.

이러한 구조적 분해를 바탕으로 한국어 특화 텍스트 처리 파이프라인은 입력 분류, 텍스트 정규화, 형태소 분석, 언어별 G2P(Grapheme to Phoneme) 처리[5]로 단계적으로 구성된다. 이 과정에서 음향 모델과 보코더 아키텍처는 기존 다국어 TTS 시스템을 유지하되, 한국어의 자음·모음 조합 체계 및 숫자·특수문자 처리 특성을 반영하여 입력 기호 정의 범위를 재정의함으로써 보다 효율적이고 언어 특화적인 텍스트 표현을 가능하게 하였다.

다국어 TTS 시스템 기반 한국어 음성 합성에서 텍스트 처리 파이프라인은 단순한 전처리 모듈이 아니라 언어 간 차이를 흡수하고 합성 안정성을 확보하는 핵심 구조로 기능한다. 본 논문에서 제시한 구조적 분해는 한국어 환경에서 텍스트 해석 단계가 수행해야 할 역할을 명확히 하며, 향후 한국어 TTS 및 에이전트 기반 음성 인터페이스 설계 시 고려해야 할 텍스트 처리 관점을 제시한다.

Ⅲ. 결론

본 논문에서는 영어 중심으로 설계된 다국어 TTS 시스템을 한국어 음성 합성에 적용할 때 발생하는 텍스트 처리 단계의 구조적 한계를 분석하고, 한국어의 언어적 특성을 고려한 텍스트 처리 파이프라인 구성 요소를 정리하였다. 숫자 및 단위 표현, 조사와 어미 결합, 혼합 언어 및 영어 약어 처리와 같이 발음 결정이 문맥에 의존하는 요소들이 텍스트 해석 단계에서 핵심적인 문제로 작용함을 논의하였으며, 이를 통해 텍스트 처리 파이프라인이 단순한 전처리를 넘어 언어 간 차이를 흡수하고 합성 안정성을 확보하는 핵심 구조로 기능함을 확인하였다. 또한 음향 모델과 보코더의 구조를 유지한 상태에서도, 텍스트 해석 단계와 입력 표현 정의를 한국어 특성에 맞게 설계함으로써 한국어 음성 합성에 필요한 언어적 정보를 효과적으로 보완할 수 있음을 보였다.

이러한 고찰은 다국어 TTS 시스템의 한국어 적용 과정에서 모델 구조 변경이나 대규모 추가 학습 없이도 텍스트 처리 단계의 설계만으로 언어 적응을 체계적으로 수행할 수 있음을 시사한다. 나아가 한국어 TTS

시스템이나 에이전트 기반 음성 인터페이스에서 다양한 입력 형태를 안정적으로 처리할 수 있는 실질적인 설계 기반을 제공한다.

참 고 문 헌

[1] Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." International Conference on Machine Learning. PMLR, 2021.

[2] 박은정, and 조성준. "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지." 제 26 회 한글 및 한국어 정보처리 학술대회 논문집 (2014): 133-136.

[3] Choi, HyunJung, et al. "Spoken-to-written text conversion for enhancement of Korean-English readability and machine translation." ETRI Journal 46.1 (2024): 127-136.

[4] Jiang, Ziyue, et al. "Dict-tts: Learning to pronounce with prior dictionary knowledge for text-to-speech." Advances in Neural Information Processing Systems 35 (2022): 11960-11974.

[5] Kyubyong Park. g2pk. https://github.com/Kyubyong/g2pk, 2019.