인공지능융합기술연구 Vol. 3, No. 1 : 1~10 May. 2025

구음장애인의 의사소통 지원을 위한 발화 재구성 시스템용 언어 모델 선정 평가

김동현[1], 유대영[1], 정현호[1], 신방호[1], 김한섭[2], 김웅식[2*]
[1]건양대학교 의료인공지능학과, [2*]건양대학교 인공지능학과

Language Model Selection Evaluation for Speech Reconstruction Systems Supporting Communication of Individuals with Dysarthria

DongHyeon Kim[1], DaeYoung Yoo[1], HyeonHo Jung[1], BangHo Shin[1], HanSeob Kim[2], WoongSik Kim[2*]
[1]Department of Medical Artificial intelligence, Konyang University
[2*]Department Artificial intelligence, Konyang University

요 약 본 연구는 구음 장애인의 음성 데이터를 활용하여 비표준 발화를 자연스러운 문장으로 변환하는 다양한 모델들의 성능을 비교 분석하고, 구음 장애인 발화 재구성에 가장 적합한 모델을 제안한다. Whisper 모델을 통해 음성에서 텍스트를 추출하고, Ko-BART와 Ko-T5 기반의 대형 언어 모델을 활용하여 문장 정제 및 자연화를 수행하였다. 두 모델은 전반적으로 유사한 성능을 보였으나, BLUERT와 BERTScore 지표에서는 소폭의 차이를 나타내어 지표 특성에 따른 모델의 상대적 강점을 시사하였다. 향후 연구에서는 발달장애인의 발화 특성을 반영하여 의사소통 효율성을 증진시키고, 일상적·사회적 상호작용에서의 장벽을 낮추는 데 기여할 수 있을 것으로 기대된다.

• 주제어 : Ko-BART, Ko-T5, BLUERT, BERTScore, 구음장애, 대규모 언어 모델

Abstract This study compares and analyzes the performance of various models that transform non-standard speech of individuals with dysarthria into natural sentences, and proposes the most suitable model for reconstructing their utterances. Speech-to-text conversion was performed using the Whisper model, and sentence refinement and naturalization were carried out using large language models based on Ko-BART and Ko-T5. While both models showed generally similar performance, slight differences were observed in the BLUERT and BERTScore metrics, suggesting relative strengths depending on the characteristics of the evaluation indicators. Future research is expected to incorporate the speech characteristics of individuals with developmental disabilities to enhance communication efficiency and reduce barriers in daily and social interactions

• Key Words : Ko-BART, Ko-T5, BLUERT, BERTScore, dysarthria, LLM

Received 29 November 2020, Revised 29 December 2020, Accepted 21 January 2023 (출판사에서 작성)
* **Corresponding Author** Woong-Sik Kim, Dept. of Artificial Intelligence, Konyang University, 158, Gwanjeodong-ro, Seo-gu, Daejeon, Korea. E-mail: wskim@konyang.ac.kr

Ⅰ. 서론

구음장애는 언어 장애의 한 형태로, 발성에 관여하는 근육의 손상이나 조절 이상으로 인해 발음이 부정확하거나 어려워지는 특성을 지닌다. 주로, 뇌성마비, 뇌졸중, 파킨슨병 등의 신경학적 질환에 의해 발생하며, 음소 왜곡, 불명확한 발화 등 효과적인 의사소통을 방해한다. 중증의 경우 사회적 고립, 정서적 문제, 교육 및 정보 접근의 제한 등 다양한 부수적 문제가 발생할 수 있다 [1]. 이러한 문제로 사회적 관계 형성, 교육, 고용 등 다양한 삶의 영역에서도 어려움을 겪고 있으며, 이를 해소하기 위한 효과적인 의사소통 지원 방안이 요구되고 있다.

최근 대규모 언어 모델(Large Language Model; LLM)의 발전은 자연어 이해 및 생성 능력을 비약적으로 향상시켰다. LLM은 대량의 텍스트 데이터를 학습하여 문맥, 의미, 문법 구조를 이해하고 자연스럽고 일관된 문장을 생성할 수 있는 기술로, 다양한 의사소통 보조 분야에서 그 가능성이 주목받고 있다 [2].

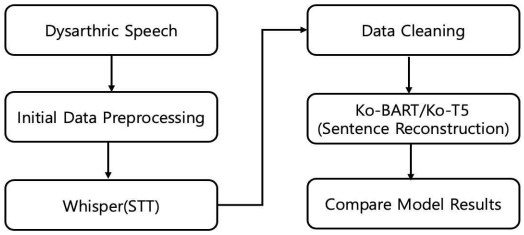

Fig. 1. Dysarthric Speech Process Pipeline.

Fig. 1은 본 연구에서 제안하는 시스템의 전체 처리 과정을 나타낸 파이프라인이다. 한국어 기반 모델인 Ko-BART와 Ko-T5를 대상으로 발화 재구성 성능을 비교·평가하였으며, 이 결과를 바탕으로 본 논문에서는 향후 시스템 개발에 필요한 언어 모델 선정의 방향성을 제시하고자 한다.

Ⅱ. 구음장애 데이터 수집 및 전처리 과정

본 연구는 “AI Hub”의 “구음장애 음성인식 데이터”를 활용하였다. 구음장애를 가진 약 1,200명의 화자가 참여하였으며, 총 5,000~5,250시간 분량의 발화와 이에 대응하는 텍스트 레이블이 JSON 파일 형태로 제공된다.

전체 데이터셋은 뇌신경 장애, 후두 장애, 언어 청각장애로 구성되어 있다. 본 연구에서는 장애 유형 간 음향적 이질성을 최소화하고, 모델 학습의 안정성을 확보하기 위해, 뇌신경 장애 데이터를 선정하여 사용하였다.

원시 음성 데이터는 무음 구간과 배경 잡음 등이 포함되어 있어, 신뢰성 높은 학습 데이터를 확보하기 위해 Fig. 2과 같은 데이터 정제 과정을 수행하였다.

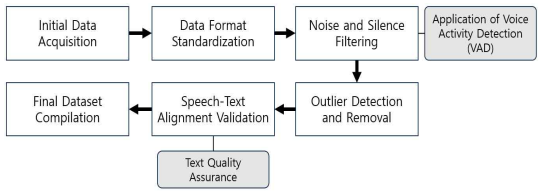

Fig. 2. Data Cleaning Process.

2.1 Data preprocessing

음성 데이터에 포함된 불필요한 정적 구간과 잡음을 제거하기 위해 Voice Activity Detection (VAD) 기법 [3]을 적용하였다. VAD는 입력된 오디오 신호의 에너지, 주파수 대역, 음성-무음 특징 등을 분석하여, 음성이 포함된 구간과 무음 또는 잡음 구간을 구분하는 알고리즘이다. 시간-프레임 단위의 short-time energy (STE)와 spectral centroid를 조합하여 각 프레임의 음성 유무를 판단하고, 이를 기반으로 무음 구간을 제거하였다. STE는 다음과 같이 정의된다.

$$E(n) = \sum_{m=0}^{M-1} x^2(n+m)$$

여기서 $x(n)$은 시간 n에서의 입력신호, M은 프레임 길이를 나타낸다. 이 값이 사전 정의된 임계값 θ를 초과할 경우 음성 구간으로 간주된다. 또한 spectral centroid $C(n)$는 다음과 같이 계산된다.

$$C(n) = \frac{\sum_{k=0}^{K=1} f(k) \cdot |X_n(k)|}{\sum_{k=0}^{K=1} |X_n(k)|}$$

여기서 $f(x)$는 주파수 $bin\,k, X_n(k)$는 n번째 프레임의 푸리에 변환 계수이다. 이는 스펙트럼의 무게 중심을 나타내면, 고주파 잡음이나 비음성 신호 제거에 활용된다. 이러한 방식으로 정제된 음성 데이터에 대해 대응하는 텍스트 레이블(JSON 파일)의 발화 길이 정보를 수정 및 동기화하여, 음성과 텍스트 간의 불일치를 최소화하였다. LLM 모델 학습을 위한 텍스트 데이터를 생성하기 위해, 정제된 음성 데이터를 OpenAI의 Whisper [4]모델을 이용하여 Speech-to-Text (STT) 변환을 수행하였다. Whisper는 OpenAI에서 개발한 범용 음성 인식 모델로, Transformer 기반의 Encoder-Decoder 구조를 가진 모델이다. 이 모델은 입력 음성을 mel-spectrogram으로 변환한 뒤, Encoder를 통해 음성 특징을 추출하고, Decoder는 텍스트 시퀀스를 생성하는 방식으로 동작한다. Whisper를 이용하여 STT 변환을 통해 얻은 텍스트는 추가 정제 과정을 거쳐 품질을 개선하였다.

2.2 Data cleaning

먼저 결측치 및 이상치가 발생한 데이터를 제거했다. 이후 Whisper로부터 얻은 변환된 텍스트와 원본 레이블 문장을 각각 마침표(.) 단위로 분할한 후, 각 문장을 대응시키는 방식으로 진행하여 품질 저하를 방지하였다. 이후 CER (Character Error Rate)과 WER (Word Error Rate)은 모두 음성을 통해 문장 단위로 의미적 병합성을 유지하며 정렬 정확도를 높일 수 있었다. 문자 오류율(CER)이 0.5 이상이거나 단어 오류율(WER)이 0.7 이상이면 해당 데이터를 제외하여 전체 데이터셋의 인식 결과의 정확도를 정량적으로 평가하기 위한 지표로, 다음과 같은 수식을 기반으로 계산된다:

$$ErrorRate = \frac{S + D + I}{N}$$

여기서 S, D, I는 각각 대체 (Substitution), 삭제(Deletion), 삽입(Insertion)된 항목의 수를 의미하며, N은 정답 시퀀스의 전체 항목 수이다. 동일한 수식 구조를 따르지만 적용 단위가 다르기 때문에 각 지표는 상이한 오류율을 나타낸다.

마지막으로, LLM 학습에 적합한 입력 형식을 갖추기 위해 모델의 최대 입력 토큰 수를 고려하여 문장 길이를 조정하거나, 일정 기준 이상일 경우 문장을 분할하여 학습 안정성을 확보하였다. 본 연구에서는 Transformer 기반의 LLM 중 T5 및 BART 계열 모델을 대상으로 실험을 수행하였으며, 각 모델의 구조적 제약과 사용한 시스템 환경을 고려하여 최대 입력 토큰 수를 T5는 256, BART는 512로 설정하였다.

이와 같은 전처리 과정을 거쳐 최종적으로 총 486개의 고품질 데이터를 구축하였으며, 이를 학습 및 평가의 객관성을 확보하기 위해 8:2의 비율로 분할하여 학습용 339개, 검증용 98개로 구성하였다.

Ⅲ. 한국어 생성 모델 비교 및 선정 과정

구음장애인의 발화를 재구성하는 시스템에 사용할 한국어 생성 모델을 선정하기 위해, 대표적인 한국어 생성 모델인 Ko-BART와 Ko-T5를 비교하였다.

3.1 Ko-BART

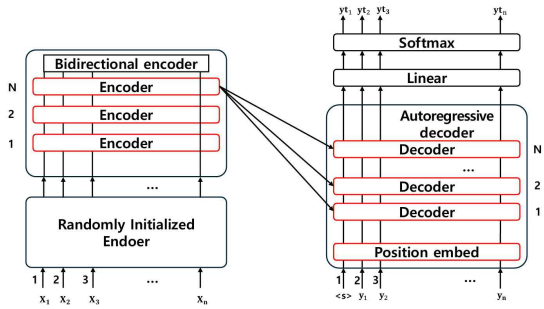

Fig. 3. Ko-BART Model.

Ko-BART는 Facebook AI에서 제안한 BART [5] (Bidirectional and Auto-Regressive Transformers) 아키텍처를 기반으로, SK텔레콤이 한국어 자연어처리 작업에 특화되도록 재학습시킨 모델이다. BART는 다양한 노이즈 삽입 기법 중 텍스트 인필링(Text Infilling) 방식을 활용하여 연속된 단어 구간을 MASK 토큰으로 치환한 뒤, 해당 구간을 원래대로 복원하도록 학습한다. 이러한 방식은 문장 내 의미 흐름을 깊이 있게 파악하고 자연스럽게 복원하는 데 효과적이며, 문법적 오류가 존재하는 입력 문장을 정제된 문장으로 변환하는 데 강점을 보인다. 생성 과정에서는 Encoder가 전체 문맥을 양방향으로 파악한 후, Decoder가 Auto-regressive 방식으로 토큰을 순차적으로 생성하며 최종 문장을 완성하게 된다. 이처럼 Ko-BART는 문장 복원과 생성 모두에 특화된 구조를 바탕으로, 구어적·비정형적 문장을 문어적·표준어 문장으로 자연스럽게 재구성할 수 있는 능력을 갖추고 있다.

3.2 Ko-T5

Ko-T5는 Google의 T5 [6] (Text-to-Text Transfer Transformer) 모델을 한국어로 확장한 모델이다. 다양한 자연어처리 작업을 Text-to-Text 형식으로 통일하여 처리하는 특징을 가진다. 입력 문장 앞에 작업 유형을 명시하는 프리픽스(prefix)를 추가함으로써, 하나의 모델로 번역, 요약, 질의응답, 문장 생성 등 여러 task를 동시에 처리할 수 있도록 설계되었다.

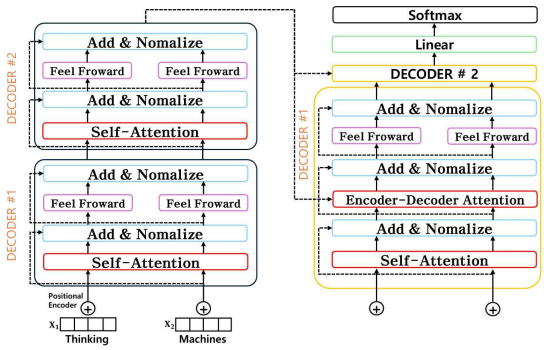

Fig. 4. Ko-T5 Model.

T5의 기본 구조는 Transformer 기반의 Encoder-Decoder 아키텍처이며, Encoder는 전체 입력 문장을 인코딩하여 문맥 정보를 요약하고, Decoder는 해당 정보를 바탕으로 출력 문장을 Auto-regressive 방식으로 생성한다. 텍스트 생성 과정에서는 입력에 태스크 지시어를 포함한 문장이 Encoder에 입력되고, Decoder는 이를 기반으로 문장을 한 단어씩 순차적으로 생성한다. 이 과정은 문맥 정보와 태스크 지시어를 함께 고려하여 적절한 문장을 생성하도록 유도하며, 다양한 작업 간 전환이 용이하다는 장점이 있다. T5는 요약, 질의응답, 문장 유사도 판단 등 범용성과 확장성 측면에서 강점을 가진 모델로 평가된다.

Ⅳ. 실험 결과

두 모델 모두 전처리된 데이터를 기반으로 파인튜닝(fine-tuning)을 수행하였으며, 특정 조건에서 모델 성능 차이를 분석하기 위해 ablation study를 진행하였다. 성능 평가는 구음장애인의 발화 데이터를 STT 변환한 후 생성된 문장을 대상으로 실시하였다.

4.1 LLM Evaluation Metrics

평가 지표로는 Rouge-L (Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence) [7]과 BLEURT [8], 그리고 BERTScore (Bidirectional Encoder Representations from Transformers Score) [9]를 사용하였다. Rouge-L은 참조 문장과 생성 문장 간의 LCS를 기반으로 계산되는 지표이다. 두 문장 X와 Y의 LCS 길이를 LCS(X, Y)라 할 때, 정밀도(precision:P), 재현율(recall:R), F-스코어(F1-score:F)는 다음과 같이 정의된다.

$$P_{LCS} = \frac{LCS(X,Y)}{|Yvert}$$

$$R_{LCS} = \frac{LCS(X,Y)}{|X|}$$

$$F_{LCS} = \frac{(1+\beta^2 P_{LCS} R_{LCS})}{\beta^2 P_{LCS} + R_{LCS}}$$

BLEURT는 BERT 기반의 사전 학습된 회귀 모델을 활용하여, 생성 문장과 참조 문장의 의미적 유사성을 예측 점수로 산출하는 학습 기반 평가 지표이다. 특히, 문법적 정확성, 자연스러움 등에 민감하게 반응하며, 사람 평가와의 상관관계가 높다.

BERTScore는 생성 문장과 참조 문장의 각 토큰 간 임베딩 기반 cosine similarity를 계산하여 정밀도, 재현율, F1-score를 산출한다. 각각은 다음과 같은 수식으로 정의된다. BLEURT와 BERTScore는 모두 의미 기반 평가 지표로 활용되나, BLEURT는 문장 품질의 세부적인 측면을, BERTScore는 표면적 일치보다는 문장 의미 보존 능력에 초점을 둔다는 점에서 유용하다.

$$P_{BERT} = \frac{1}{|c|}\sum_{j=1}^{|c|} \max_i \cos(e_{c_j}, e_{r_i})$$

$$R_{BERT} = \frac{1}{|r|}\sum_{i=1}^{|r|} \max_j \cos(e_{r_i}, e_{c_j})$$

$$F_{BERT} = \frac{2 \cdot P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}}$$

Table. 1. Model Performance Comparison

Evaluatuin Metric	Ko-BART	Ko-T5
Rouge-L	0.855	0.886
BLEURT	0.787	0.765
BERTScore	0.9358	0.963

실험 결과, 두 모델 모두 전반적으로 높은 성능을 보였으나, 지표별로 소폭의 차이가 나타났다. Rouge-L과 BERTScore에서는 Ko-T5가 우수한 성능을 보였고, BLEURT에서는 Ko-BART가 상대적으로 뛰어난 결과를 기록하였다. 이는 평가 지표의 특성에 따라 모델의 상대적 강점이 달라질 수 있음을 시사한다. 본 연구의 목적은 구음장애인의 발화 재구성 시스템을 개발하는 것으로, 원문의 의미와 뉘앙스를 정확히 보존하는 것이 핵심 과제이다. 그러므로 의미적 정밀도에서 강점을 보인 Ko-BART 모델이 적합할 것으로 판단된다.

Ⅴ. 결론

본 연구는 구음장애인의 의사소통 지원을 위한 발화 재구성 시스템 개발의 적절한 한국어 언어 생성 모델을 선정하기 위해 Ko-BART와 Ko-T5 모델을 비교・분석하였다. 실험 결과, 의미 보존 및 문맥 정밀도 측면에서 우수한 성능을 보인 Ko-BART 모델이 본 시스템에 적합함을 확인하였다. 향후 연구에서는 Ko-BART 모델을 중심으로 한 실시간 발화 재구성 시스템을 개발하고자 한다.

학습 손실 추이 분석과 다양한 외부 도메인 데이터셋을 활용한 추가 실험을 통해 모델의 일반화 가능성을 평가하고, 구음장애인의 다양한 발화 특성을 반영하는 방향으로 모델을 개선하고자 한다. 이를 통해 구음장애인의 실질적인 의사소통 효율성을 향상시키고, 일상생활 및 사회적 상호작용에서 겪는 장벽을 실질적으로 완화시킬 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학사업 지원을 받아 수행되었음 (2024-0-00047). 또한, 대학혁신지원사업의 캡스톤디자인 교과목 운영 지원을 받아 수행되었음. 이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "구음장애 음성인식 데이터"를 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

REFERENCES

[1] Enderby, Pam. "Disorders of communication: dysarthria." Handbook of clinical neurology 110 (2013): 273-281.

[2] Hadi, Muhammad Usman, et al. "A survey on large language models: Applications, challenges, limitations, and practical usage." Authorea Preprints 3 (2023).

[3] Patil, Rajesh Maharudra, and C. M. Patil. "Unveiling the State-of-the-Art: A Comprehensive Survey on Voice Activity Detection Techniques." 2024 Asia Pacific Conference on Innovation in Technology. IEEE

[4] Vestman, V., Gowda, D., Sahidullah, M., Alku, P., & Kinnunen, T. (2018). Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction. Speech Communication, 99, 62-79.

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

[6] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research

[7] Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. Information Sciences Institute, University of Southern California

[8] Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning Robust Metrics for Text Generation. Google Research

[9] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. ICLR 2020

저자소개

김 동 현 (Dong-Hyeon Kim)
2020년 3월~현재 : 건양대학교 인공지능학과 학사과정
관심분야 : LLM, Multimodal, Vision, Efficient Deep Learning Architectures, NLP

유 대 영 (Dae-Young Yoo)
2020년 3월~현재 : 건양대학교 인공지능학과 학사과정
관심분야 : VLM, LLM, Robotics, Vision, Self-Driving

정 현 호 (Hyeon-Ho Jung)
2020년 3월~현재 : 건양대학교 인공지능학과 학사과정
관심분야 : VLM, LLM, Vision, NLP, Multimodal

신 방 호 (Bang-Ho Shin)
2021년 3월~현재 : 건양대학교 인공지능학과 학사과정
관심분야 : 인공지능, 머신러닝, 빅데이터, LLM, NLP

김 한 섭 (Han-Seob Kim)
2019.02 조선대학교 컴퓨터공학과 (공학사)
2021.08 고려대학교 컴퓨터학과 (공학석사)
2025.02 고려대학교 컴퓨터학과 (공학박사)
2019.01~2024.01 한국과학기술연구원 인공지능연구단 인턴연구원
2025.03~현재 건양대학교 인공지능학과 조교수
관심분야 가상현실, 증강현실, HCI, 감성컴퓨팅, 가상인간, 인공지능

김 웅 식 (Woong-Sik Kim)
1989년 2월 : 인하대학교 정보공학과(공학석사)
2007년 2월 : 인하대학교 컴퓨터공학과(공학박사)
2006년 3월~현재 : 건양대학교 인공지능학과 교수
관심분야 : 인공지능, 의료공학, 임베디드, 뇌파