대규모 언어 모델 기반 한국어 휴지 예측 연구

나정호°[1], 이정[1], 나승훈[1], 정정범[2], 최맹식[2], 이충희[2]
[1]전북대학교, [2](주)엔씨소프트
{jhna2023, fhqlatm, nash}@jbnu.ac.kr, {jbjeong, mschoi, forever73}@ncsoft.com

A Study on Korean Pause Prediction based Large Language Model

Jeongho Na°[1], Joung Lee[1], Seung-Hoon Na[1], Jeongbeom Jeong[2], Maengsik Choi[2], Chunghee Lee[2]
[1]Jeonbuk National University, [2]NCSOFT Corp.

요 약

본 연구는 한국어 음성-텍스트 데이터에서 보편적으로 나타난 휴지의 실현 양상을 분석하고, 이를 토대로 데이터셋을 선별해 보편적이고 규격화된 한국어 휴지 예측을 위한 모델을 제안하였다. 이를 위해 전문적인 발성 훈련을 받은 성우 등의 발화가 녹음된 음성-텍스트 데이터셋을 수집하고 MFA와 같은 음소 정렬기를 사용해 휴지를 라벨링하는 등의 전처리를 하고, 다양한 화자의 발화에서 공통적으로 나타난 휴지를 선별해 학습데이터셋을 구축하였다. 구축된 데이터셋을 바탕으로 LLM 중 하나인 KULLM 모델을 미세 조정하고 제안한 모델의 휴지 예측 성능을 평가하였다.

주제어: 휴지, 휴지 예측, 끊어 읽기, LLM, Large Language Model

1. 서론

최근 AI(Artificial intelligence)가 탑재된 스마트 스피커의 수요가 증가하면서 자연스러운 음성을 합성하는 TTS(Text to Speech) 기술에 대한 관심 역시 높아지고 있다. TTS에서 자연스러운 음성을 합성하기 위해서는 대규모의 고품질 음성 데이터뿐만 아니라 조음 활동에서 나타나는 휴지(Pause)에 대한 분석 역시 요구된다. 표준국어대사전에서 나타난 휴지의 사전적 정의는 '조음 활동의 일시적인 정지를 뜻하는 것'으로 주로 단어와 단어, 어절과 어절, 문장과 문장 사이에 사용되는 것으로 알려져 있다[1]. 이러한 휴지는 조음 활동에서 호흡, 발화의 강조나 기획, 주목을 끄는 등의 역할을 수행하며, 청자가 화자의 발화를 이해하고 처리하기 위한 시간을 제공한다[1][2]. 이와 같이 휴지는 발화에서 다양한 역할을 수행하고 있어 자연스러운 음성합성을 위해 고려해야 할 필수적인 요소이다.

그러나 휴지는 화자의 억양이나 습관, 사회적 배경 등에 따라 실현 양상에 차이를 보이기 때문에 단일 언어에서 보편적으로 나타나는 휴지의 특성에 대한 분석이 필요하다. 하지만 휴지 예측과 관련된 대다수의 선행 연구에서는 음성 데이터에 나타나는 휴지의 실현 양상을 고려하지 않고 구문 특징이나 언어적 표현만을 사용하거나 화자 임베딩과 같은 화자의 발화 특징만을 사용하고 있었다[3][4][5]. 이처럼 시간이나 비용 등과 같은 현실적인 이유로 여러 명의 화자가 포함된 다중화자 데이터셋을 사용하는 현 실정에서 보편적인 휴지의 특성을 반영하지 않을 경우 다양한 화자의 발화 특징이 뒤섞여 오히려 모델의 휴지 예측 성능이 감소할 것으로 예상된다.

또한, 한국어에 나타는 언어적 표현이나 구문 특징, 독특한 운율적 특성을 모델이 학습하기 위해서는 방대한 규모의 데이터셋이 필요하다. 그러나 고품질의 데이터셋을 구축하는 것은 현실적인 어려움이 존재하며, 데이터셋에 숨어있는 규칙을 찾기 위해선 대규모의 매개변수를 이용한 통계 모델링이 요구된다. 따라서 최근 자연어 처리 분야에서 두각을 보이는 대규모 언어 모델(Large Language Model, LLM)을 활용하면 휴지 예측 분야에서 유의미한 성능을 보일 수 있을 것으로 추정된다. 더구나 LLM은 적절한 프롬프트를 사용하면 적은 양의 데이터만을 가지고 특정 사용 사례(use-cases)에 맞게 모델을 조정할 수 있어 이후 다양한 화자에게서 나타나는 개별적인 발화 특징을 모델에 학습시키기 용이할 것으로 보인다[6].

이에 본 연구에서는 다양한 화자의 발화 특징을 반영한 한국어 휴지 예측 모델의 사전 연구로 LLM에 기반한 보편적이고 규격화된 휴지 예측 모델을 제안하고자 한다. 이를 위해 첫째, 한국어 음성-텍스트 데이터를 수집하고 음소 정렬(Forced Align)을 통해 휴지를 라벨링하였다. 둘째, 구축한 데이터셋을 분석하여 여러 화자의 발화에서 공통적으로 나타난 휴지를 분석하고 데이터를 선별해 학습데이터와 평가데이터를 구축하였다. 마지막으로 정제된 데이터를 바탕으로 Polyglot 모델을 기반으로 한 KULLM 모델[2]을 파인튜닝(Fine-tuning)하고 휴지 예측에 대한 모델의 성능을 평가하였다.

2. 관련 연구

2.1 한국어 휴지 연구

안병섭(2007)은 남여 아나운서가 낭독한 음성 데이터를 분석하여 발화 단위 내부에서 나타나는 휴지의 실현 양상을 살펴본

[1]https://stdict.korean.go.kr/search/searchView.do

[2]https://github.com/nlpai-lab/kullm

후, 그 결과를 토대로 선행 연구에서 분석한 휴지의 역할에 대해 반성적 검토를 하였다. 분석 결과, 음성 데이터에서 하나 이상의 음운론적 단어로 구성된 강세구가 하나 이상의 강세구로 형성된 억양구에 비해 2배 가까이 많았지만, 강세구 경계에서 휴지가 나타난 비율이 0.25%, 0.40%인 반면 억양구 경계에서 휴지가 실현된 비율이 96.2%, 89.3%로 나타나 휴지는 일반적으로 억양구 경계에서 실현된다고 주장하였다[7]. 또한, 신지영(2011)은 일반인의 자유 발화 및 낭송 발화를 분석하여 한국어에서 나타나는 음운 규칙을 살펴보았다. 그 결과, 강세구에서는 약한 휴지가 발견되며 억양구에서는 억양구의 마지막 음절에서 관찰되는 특징적인 음높이 패턴과 어말 자음화, 물리적인 휴지의 실현으로 청자로 하여금 음운구에 비해 큰 휴지를 느끼게 한다고 하였다[8]. 이처럼 음운 규칙은 연구자나 연구에 사용한 음성 데이터에 따라 서로 다른 결과를 보여 휴지 예측을 위한 표준화된 규칙을 정의하기는 어려울 것으로 보인다.

2.2 LLM 기반 휴지 예측 연구

Wang et al.(2023)은 PLM(Pretrained Language Model)과 LLM을 사용해 ESL(English as a Second Language) 학습자의 발화에서 끊어 읽기를 평가하는 방법을 제안하였다. 실험 결과, PLM을 사용하면 라벨링된 학습데이터에 대한 의존도가 크게 감소하고 성능이 향상되는 것을 확인하였으며, LLM인 ChatGPT가 끊어 읽기 분야에서 더 발전할 수 있는 잠재력을 가지고 있음을 확인하였다[9].

3. 데이터셋

3.1 데이터셋 수집

표 1. 한국어 음성-텍스트 데이터 분포

데이터명	발화문	휴지	화자
감성 및 발화스타일	632,105	574,376	49
문학작품 낭송	208,234	478,003	46

한국어에서 나타나는 보편적인 휴지의 예측을 위해 다음과 같은 데이터 수집 및 전처리 과정을 수행하였다. 첫째, Ai-hub[3], Kaggle[4] 등의 포털에서 한국어 화자 음성-텍스트 데이터를 수집하였다. 이때 일반인으로 녹음한 데이터나 자유발화 데이터의 경우 화자 개개인의 억양이나 습관, 상황이나 맥락 등의 발화 특징으로 인해 휴지의 실현 양상이 크게 차이나기 때문에 1) 전문적인 발성 훈련을 받은 성우가 녹음한 데이터와 2) 동일한 대본을 다양한 화자가 발화한 데이터를 우선적으로 수집하였다.

[3]https://www.aihub.or.kr/
[4]https://www.kaggle.com/

둘째, Montreal Forced Aligner(MFA)[10]를 사용하여 발화문에서 휴지를 라벨링하였다. MFA는 단어가 전환될 때 발생하는 30ms 이상의 소리 없는 구간을 휴지로 인식한다. 이중 발화문의 처음과 끝에 등장하는 휴지는 음성녹음 과정에서 발생한 지연으로 간주하고 제거했으며, 나머지를 SP(silent Pause)로 라벨링하였다.

이러한 데이터 수집 및 라벨링을 통해 구축한 데이터는 표 1과 같다. '감성 및 발화스타일별 음성합성 데이터(감성 및 발화스타일)'는 49명의 화자가 발화한 632,105 건의 데이터로 총 574,376개의 휴지를 라벨링하였다. 또한, '문학작품 낭송, 낭독 데이터(문학작품 낭송)'는 46명의 화자가 발화한 208,234 건의 데이터로 총 478,003개의 휴지가 라벨링되어있다. 해당 데이터는 모두 동일한 대본을 다양한 성우가 발화한 음성을 녹음한 데이터로 데이터 선별을 통해 학습데이터를 구축하였다.

3.2 데이터셋 선별

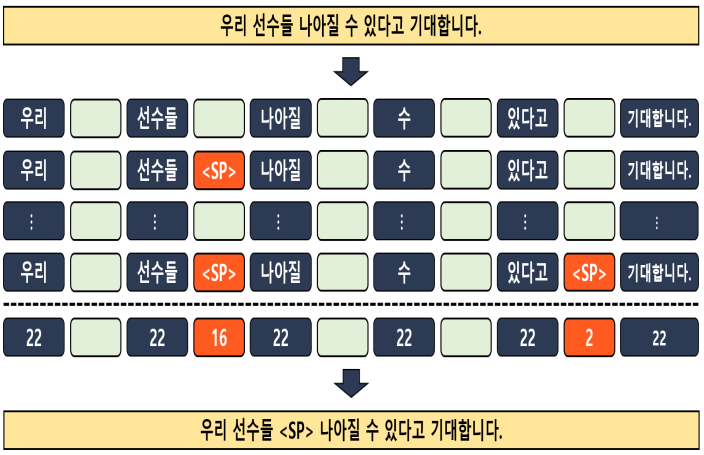

그림 1. 학습데이터 선별 방법

그림 2. 학습데이터 예시

표 2. 학습데이터 분포

Data	Threshold	Train		Validation		Test	
		Sentence	Pause	Sentence	Pause	Sentence	Pause
감성 및 발화스타일 + 문학작품 낭송	(휴지 등장 횟수/화자 수) ≠ 0.0	40,000	126,819	5,000	15,708	5,000	15,852
	(휴지 등장 횟수/화자 수) ≥ 0.1	40,000	114,511	5,000	14,174	5,000	14,341
	(휴지 등장 횟수/화자 수) ≥ 0.2	40,000	93,918	5,000	11,647	5,000	14,341
	(휴지 등장 횟수/화자 수) ≥ 0.3	40,000	80,524	5,000	9,910	5,000	9,966
	(휴지 등장 횟수/화자 수) ≥ 0.4	40,000	71,981	5,000	8,900	5,000	8,912
	(휴지 등장 횟수/화자 수) ≥ 0.5	40,000	64,985	5,000	7,958	5,000	8,025
	(휴지 등장 횟수/화자 수) ≥ 0.6	40,000	55,587	5,000	6,750	5,000	6,834
	(휴지 등장 횟수/화자 수) ≥ 0.7	40,000	48,046	5,000	5,863	5,000	5,911
	(휴지 등장 횟수/화자 수) ≥ 0.8	40,000	40,810	5,000	4,956	5,000	5,050
	(휴지 등장 횟수/화자 수) ≥ 0.9	40,000	30,435	5,000	3,700	5,000	3,809
	(휴지 등장 횟수/화자 수) = 1.0	40,000	26,626	5,000	3,263	5,000	3,324

본 연구에서 제안하는 보편적이고 규격화된 휴지 예측 모델을 위해 그림 1과 같이 데이터를 선별하여 학습데이터를 구축하였다. 데이터 선별 조건은 첫째, 최소 3명 이상의 화자가 발화한 대본을 추출하였다. 이는 수집한 데이터셋이 최소 1명에서 최대 65명이 동일한 대본을 사용해 녹음을 했기 때문으로 여러 화자에게서 나타나는 휴지를 살펴보기 위해 기준점을 정하고 데이터를 전처리하였다. 둘째, 대본별로 휴지를 라벨링한 발화문을 띄어쓰기를 기준으로 토큰화한 뒤 발화문에 나타난 휴지의 위치와 등장 횟수를 확인하였다. 마지막으로 각 위치별 휴지의 등장 횟수를 해당 대본을 발화한 화자의 수로 나누고 임계점(Threshold)별로 분류해 그림 2와 같이 학습데이터를 구축하였다. 임계점은 0.0에서 1.0까지 총 11단계로 구분하고 이후의 연구를 진행하였다.

이와 같은 과정을 통해 총 5만 건의 데이터셋을 선별했으며, 전체 데이터셋을 학습 및 검증, 평가 데이터셋으로 각각 8:1:1 비율로 나뉘어 모델 학습을 진행하였다. 최종적으로 모델 학습에 사용한 데이터셋의 분포는 표 2와 같다.

4. 실험 설계

한국어에서 보편적으로 나타나는 휴지를 예측하기 위해 비영어권 언어에서 높은 성능을 보이는 Polyglot 모델을 기반으로 한 KULLM-Polyglot-5.8B-v2 모델을 그림 3과 같이 프롬프트를 구성해 미세 조정(Fine tuning)하였다. 프롬프트는 모델이 수행할 작업을 정의한 'Instruction', 모델의 응답의 질을 높이기 위해 제공하는 추가 문맥인 'Context', 응답받고자 하는 입력과 입력 받은 값에 대한 출력인 'Input'과 'Output' 4가지로 구성했으며, 'Context'에 입력한 휴지의 대한 정의는 표준국어대사전을 참조하여 구성하였다. 실험은 Tesla V100-DGXS-32GB 3기로 진행하였으며 각 GPU당 배치크기를 4*8로 하여 4 epochs 동안 학습을 수행하였다. 또한, 옵티마이저(Optimizer)는 Adafactor, 학습률(Learning Rate)는 2e-6으로 설정한 뒤 50 steps동안 warmup 후 선형적으로 감소하도록 스케줄링하였다. 그 후, 표 2의 데이터셋별로 모델을 학습한 뒤 정밀도(Precision)와 재현율(Recall), F_1 점수(F1 score)를 사용해 모델을 정량 평가하고, 모델간 성능을 비교분석하였다.

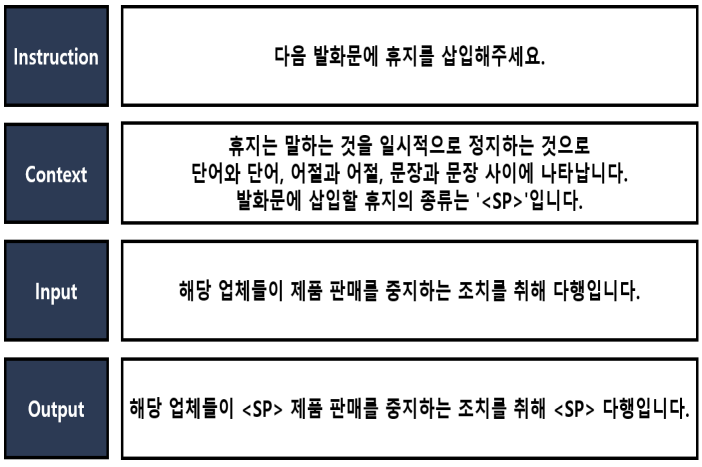

그림 3. 프롬프트 예시

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (2)$$

$$F_1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

5. 평가 결과

표 3. 한국어 휴지 예측 모델 성능

Data	Threshold	Precision	Recall	F_1 Score
감성 및 발화스타일 + 문학작품 낭송	(휴지 등장 횟수/화자 수) $\neq 0.0$	0.8217	0.8087	0.8151
	(휴지 등장 횟수/화자 수) ≥ 0.1	**0.8224**	0.8098	**0.8160**
	(휴지 등장 횟수/화자 수) ≥ 0.2	0.8182	0.8082	0.8132
	(휴지 등장 횟수/화자 수) ≥ 0.3	0.8173	**0.8144**	0.8158
	(휴지 등장 횟수/화자 수) ≥ 0.4	0.8017	0.8078	0.8048
	(휴지 등장 횟수/화자 수) ≥ 0.5	0.8032	0.8041	0.8037
	(휴지 등장 횟수/화자 수) ≥ 0.6	0.7930	0.7749	0.7838
	(휴지 등장 횟수/화자 수) ≥ 0.7	0.7820	0.7762	0.7791
	(휴지 등장 횟수/화자 수) ≥ 0.8	0.7655	0.7750	0.7702
	(휴지 등장 횟수/화자 수) ≥ 0.9	0.7503	0.7410	0.7456
	(휴지 등장 횟수/화자 수) $= 1.0$	0.7319	0.7182	0.7250

본 논문에서 제안한 모델의 한국어 휴지 예측 성능 평가 결과는 표 3과 같다. 평가 결과, 등장 빈도(휴지 등장 횟수/화자 수)가 0.1을 넘는 휴지를 선별한 데이터셋으로 학습한 모델이 F_1 Score기준 0.8160으로 가장 높은 성능을 보였으며, 등장 빈도 0.3 이상의 모델이 0.8158, 0.0 이상의 모델이 0.8151 순으로 성능이 높은 것이 확인되었다.

이는 휴지의 등장 빈도가 높아질수록 다양한 화자에게서 공통적으로 나타나는 휴지가 선별되어 모델의 성능이 높아질 것이라는 예측과는 상반된 결과이다. 그 원인으로는 첫째, 수집한 데이터셋이 기쁨, 슬픔, 분노 등의 감정을 포함하고 있기 때문으로 추정된다. 즉, 감성별로 나타나는 휴지의 실현 양상이 차이를 보이기 때문에 휴지의 등장 빈도가 높아질수록 휴지가 나타나는 양상이 세분화되어 모델의 성능에 영향을 준 것으로 보인다.

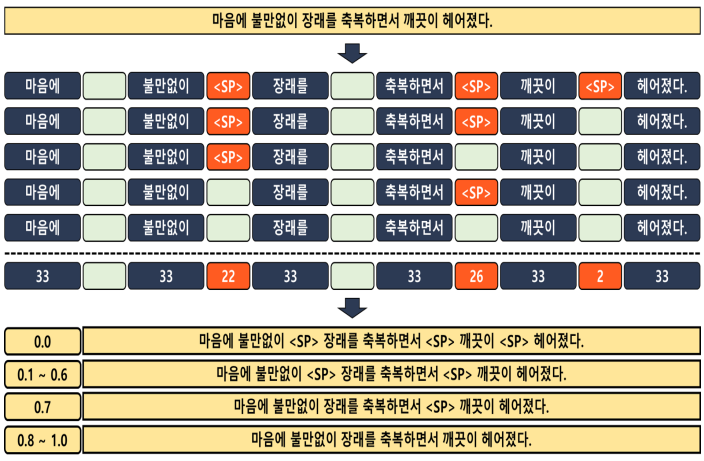

그림 4. 임계점별 데이터 선별 예시

둘째, 전문적인 발성 훈련을 받은 성우일지라도 연기 방식 등에 의해 끊어 읽기 방식에서 차이를 보인다. 실제로 그림 4에서 보이는 것처럼 임계점을 높일수록 선별된 휴지가 점점 줄어드는 것을 볼 수 있는데, 성우에 따라 휴지가 실현될 수 있을만큼 충분히 긴 문장에서도 쉼 없이 발화하는 경우가 있기 때문이다. 이로 인해 휴지의 등장 빈도가 높아질수록 휴지의 실현 양상이 보편적이지 않게 되면서 정량 평가 결과가 낮게 나타난 것으로 보인다.

마지막으로 데이터셋 선별에 단순한 통계적 요인만을 사용하였기 때문이다. 본래 연구에서 제안하는 보편적이고 규격화된 휴지 예측 모델을 위해 표준어로 사용하고 있는 서울말에서 공통적으로 나타나는 휴지의 실현 양상을 조사하고 이를 기반으로 데이터셋을 선별하고자 하였다. 그러나 선행 연구에서도 분석에 사용한 음성 데이터에 따라 결과에 차이를 보이고 있었다. 비교적 최근 연구에서조차 하나의 언어에서 나타나는 음높이, 크기, 길이, 휴지 등과 같은 운율은 화자마다 다르게 나타나며 운율을 구별하기 위해서는 연구자가 운율적 요소를 바라보는 감각에 의존할 수 밖에 없다고 말하고 있었다[11]. 따라서 보다 정확한 모델의 성능 평가를 위해서는 모델의 추론 결과를 음성합성하고 합성된 음성샘플을 통해 선호도 테스트(preference test) 등과 같은 정성 평가가 이루어져야 할 것으로 보인다.

6. 결론

본 논문에서는 한국어에서 보편적으로 나타나는 규격화된 휴지 예측을 위해 데이터 분석과 선행 연구에서 나타는 휴지의 특성을 반영하여 데이터셋을 선별하고 학습을 진행하였다. 학습을 위해 최근 자연어 처리 분야에서 각광을 받고 있는 LLM을 사용하고자 Polyglot 모델을 기반으로 한 KULLM 모델을 사용하였으며, 프롬프트를 구성해 주어진 문장에서 휴지의 위치를 예측하는 모델을 설계하였다. 연구 결과, 제안된 모델이 전반적으로 높은 성능을 보여 휴지 예측 연구에서 데이터의 선별과 LLM의 활용 가치를 입증하였다.

그러나 관련 연구에서조차 표준화된 휴지 실현 양상을 제시하지 못하고 있었으며, 본 연구에서도 데이터셋 선별에 통계적인 요인만을 사용하여 학습데이터에 보편적으로 나타나지 않는 휴지 실현 양상이 포함되어 정성 평가만으로는 정확한 모델의 성능을 평가하는데 한계가 있다. 이에 정량 평가를 실시해 LLM이 예측한 휴지의 위치가 자연스러운지에 대한 평가가 필요하다. 향후 연구에서는 모델의 추론 결과를 음성합성하고 합성된 음성샘플을 통해 정성 평가를 수행해 보편적으로 나타나는 휴지를 예측하는 모델을 선정하고자 한다.

참고문헌

[1] D. Yoo and J. Shin, "A realization of pauses in utterance across speech style, gender, and generation," *Phonetics and Speech Sciences*, Vol. 11, No. 2, pp. 33–44, 2019.

[2] J. Shin, "Breath and memory in speech based on quantitative analysis of breath groups and pause units in korean," *Korean Linguistics*, Vol. 79, pp. 91–116, 2018.

[3] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in japanese text-to-speech synthesis," *arXiv preprint arXiv:2104.12395*, 2021.

[4] Y. Matsunaga, T. Saeki, S. Takamichi, and H. Saruwatari, "Spontaneous speech synthesis with linguistic-speech consistency training using pseudo-filled pauses," *arXiv preprint arXiv:2210.09815*, 2022.

[5] D. Yang, T. Koriyama, Y. Saito, T. Saeki, D. Xin, and H. Saruwatari, "Duration-aware pause insertion using pre-trained language model for multi-speaker text-to-speech," *arXiv preprint arXiv:2302.13652*, 2023.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[7] B. Ahn, "A critical review of the studies on the functions of pause in korean," *The Studies of Korean Language and Literature*, No. 28, pp. 67–87, 2007.

[8] J. Shin, *Korean Phonetics and phonology.* Pagijong, 2011.

[9] Z. Wang, S. Mao, W. Wu, Y. Xia, Y. Deng, and J. Tien, "Assessing phrase break of esl speech with pre-trained language models and large language models," *arXiv preprint arXiv:2306.04980*, 2023.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." *Interspeech*, Vol. 2017, pp. 498–502, 2017.

[11] J. Oh, "Distinguishing features and variability of intonation patterns in korean phonological phrases: The effect of morpheme boundaries," *The Journal of Yeongju Language Literature*, Vol. 52, pp. 103–138, 2022.