대화 맥락에 기반한 한국어 휴지 예측 모델

이정[1], 나정호[2], 정정범[3], 최맹식[4], 이충희[5], 나승훈[*6]
전북대학교[1,2,*6], ㈜엔씨소프트[3,4,5]
{fhqlatm, jhna2023, nash}@jbnu.ac.kr, {jbjeong, mschoi, forever73}@ncsoft.com

Korean Pause Prediction Model based on Dialogue Context

Joung Lee[1], Jeongho Na[2], Jeongbeom Jeong[4], Maengsik Choi[5], Chunghee Lee[6], Seung-Hoon Na[*6]
Jeonbuk National University[1,2,*6], NCSOFT Corp.[3,4,5]

요 약

음성 사용자 인터페이스(Voice User Interface)에 대한 수요가 증가함에 따라 음성 합성(Speech Synthesis) 시스템에서 자연스러운 음성 발화를 모방하기 위해 적절한 위치에 휴지를 삽입하는 것이 주된 과업으로 자리잡았다. 대화의 연속성을 고려했을 때, 자연스러운 음성 기반 인터페이스를 구성하기 위해서는 대화의 맥락을 이해하고 적절한 위치에 휴지를 삽입하는 것이 필수적이다. 이에 따라 본 연구는 대화 맥락에 기반하여 적절한 위치에 휴지를 삽입하는 Long-Input Transformer 기반 휴지 예측 모델을 제안하고 한국어 대화 데이터셋에서 검증한 결과를 보인다.

주제어: 음성 사용자 인터페이스, 휴지 예측, Long-Input Transformer

1. 서론

최근 다양한 산업 분야에서 인공지능(Artificial intelligence, AI)과 결합한 서비스를 앞다투어 출시하여 실생활에서 어렵지 않게 찾아볼 수 있다. 인공지능 관련 서비스는 차량 내의 인공지능 스피커부터 ChatGPT나 CLOVA-X와 같은 초거대 언어 모델(e.g. GPT-4 [1], HYPERCLOVA [2])에 기반한 대화형 서비스까지 다양한 규모에서 활용 사례를 보인다. 이처럼 다양한 인간-로봇 상호작용(Human-Robot Interaction, HRI) 시스템이 출시되면서 보다 편리한 사용자 인터페이스(User Interface, UI)에 대한 관심이 증대되고 있으며, 이에 따라 음성 언어 처리 기술에 기반한 음성 사용자 인터페이스(Voice User Interface)도 더욱 활발히 연구되고 있다 [3, 4].

음성 사용자 인터페이스에 대한 수요가 증가함에 따라 자연스레 음성 합성(Speech Synthesis) 분야에서도 대규모 사전 학습 데이터를 활용한 모델이 공개되며 큰 성능 향상을 보였다 [5, 6, 7]. 음성 합성 기술은 인간의 발화 음성에 해당하는 음파를 기계가 자동으로 합성하는 기술로 자연스러운 음성 발화를 모방하기 위해 발화의 운율이나 억양 및 휴지 등이 적절하게 삽입되어야 한다.

휴지(Pause)는 조음 활동에서 나타나는 일시적인 정지를 뜻하며 발화 내에서 의미를 확정하거나 호흡, 강조 등 다양한 기능을 수행한다 [8]. 휴지의 기능적 측면을 고려했을 때, 자연스러운 합성 음성을 생성하기 위해서는 휴지가 필수적으로 고려되어야 한다. 특히 휴지는 문장 부호에서 발생하는 휴지(Punctuation-Indicated Pause)와 발화 특징에 따라 나타나는 휴지(Respiratory Pause)로 분류되는데, 발화 특징에 따른 휴지의 경우 발화 스타일이나 배경, 습관 등에 의해 발화자에 따라 매우 다양한 분포를 보인다 [9, 10]. 하지만 전문적으로 발화 훈련을 받은 성우나 아나운서에게 동일한 대본을 제공하더라도 휴지 삽입 위치가 상이하기 때문에 적절한 위치에 휴지를 삽입하는 것이 주된 과업으로 자리 잡았다 [11, 12].

대부분의 선행 연구는 발화문에서 나타나는 언어적 표상(Representation)이나 발화자의 화자 특징 벡터(Embedding Vector)만을 활용하여 휴지의 삽입 위치를 예측한 연구 결과를 보여왔다 [13, 14]. 하지만 음성 인터페이스에 대한 사용자 경험을 폭넓게 증가시키기 위해서는 자연스러운 음성 합성 시스템과 더불어 이전 대화에 기반한 연속적인 상호 작용이 전제되어야 한다. 대화의 연속성을 고려했을 때, 자연스러운 음성 기반 인터페이스를 구성하기 위해서는 대화의 맥락을 이해하고 적절한 위치에 휴지를 삽입하는 것이 필수적이다. 이에 따라 본 연구는 대화 맥락에 기반하여 적절한 위치에 휴지를 삽입하는 Long-Input Transformer 기반 휴지 예측 모델을 제안하고 한국어 대화 데이터셋에서 검증한 결과를 보인다.

본 논문의 주요한 기여는 다음과 같다.

- 대화 맥락에 기반한 한국어 발화 내 휴지 삽입 위치 예측 모델 제안
- 음성-음소 데이터 간의 정렬을 통한 한국어 대화 데이터 재구성 및 라벨링
- Long-Input Transformer 구조 모델을 활용하여 긴 대화의 구문 표상을 획득하고 휴지 생성 과정에 활용

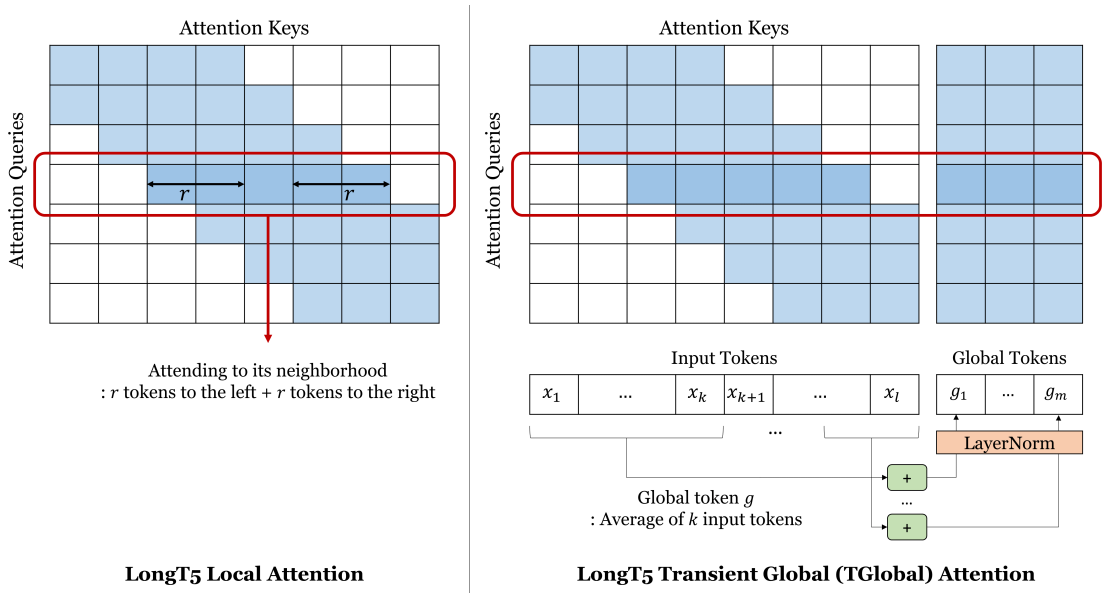

그림 1. LongT5 Attention 연산 작용 과정

2. 관련 연구

2.1 Montreal Forced Aligner

Montreal Forced Aligner (MFA) [1]는 음성 발음 사전을 활용하여 음성-음소 간의 정보를 학습하고 정렬하는 프로그램이다. MFA는 입력된 문장과 음성 데이터로부터 음소 혹은 어절 단위의 정렬 정보를 제공하며 특히 발화 음성 내 음소 간 휴지 정보 또한 제공하는 장점을 보인다 [15]. 본 연구는 음성 및 발화 대본으로 구성된 데이터셋으로부터 획득한 정렬 정보를 통해 얻은 발화 내 휴지 위치를 활용한다.

2.2 Long-Input Transformer

기존 Transformer 구조의 모델은 Attention 연산량이 입력 길이의 제곱에 비례하기 때문에 GPU 메모리 자원의 한계로 인해 입력 문장의 길이가 제한되었다. 하지만 입력 문장을 제한하는 경우 긴 문장 입력을 효과적으로 처리할 수 없기에 이를 확장하는 여러 연구가 수행되었다. ETC [16] 및 BigBird [17]은 입력 문장에 대해 임의의 글로벌 토큰 및 무작위 토큰을 지정하여 Local (Sparse) Attention 연산을 수행하면서도 멀리 떨어진 토큰과의 관계를 포착하는 기법을 제안하였다. 또한 LongT5 [18]은 기존 T5 모델을 긴 문장 입력이 가능하도록 TGlobal 토큰에 기반한 Attention 기법을 제안하였다. 이는 추가적인 입력 없이 임시로 구성된 TGlobal 토큰을 활용하여 멀리 떨어진 토큰과의 관계를 포착 가능하다는 장점이 있다.

ETC와 BigBird 및 LongT5 모두 문장 길이가 제한된 기존 모델에 비하여 긴 문맥을 활용한 과업에서 월등한 성능을 보인바 있다. 본 연구에서는 긴 대화 내에서 이전 발화를 최대한 포함하여 맥락 정보를 상세히 포착하고 휴지 예측에 활용할 수 있도록 LongT5 모델을 활용하였다.

2.3 휴지 예측 모델

[13, 14]는 발화문 내 휴지를 예측하는 과업을 각 토큰별 휴지 발생 확률에 의한 분류 문제로 정의하고 발화문의 표상을 활용하도록 사전 학습된 BERT 및 BiLSTM에 기반한 휴지 예측 모델을 제안하였다. 또한 Pausespeech [19]는 사전 학습된 언어 모델에 기반한 발화문의 구문 특성을 기존 음성 합성 프레임워크에 통합하여 휴지 예측에 활용할 수 있도록 End-to-End 방식의 음성 합성 시스템을 제안하였다.

3. 제안 방법

3.1 대화 맥락에 기반한 한국어 휴지 예측 모델

그림 1은 LongT5에서 제시한 Local 및 TGlobal Attention 연산의 작용 과정을 나타낸 그림이다. Local Attention은 각 토큰으로부터 좌우 각 r개의 토큰과의 관계를 포착한다. 또한 입력 토큰들에 대한 Layer noarmalization 연산을 통해 TGlobal(Transient Global) 토큰을 구성하여 멀리 떨어진 토큰과의 관계도 포착한다. 입력 문장의 길이 l 및 TGlobal 토큰의 수 l/k에 대하여 Local Attention의 연산 복잡도는 $O(l \times r)$이며, TGlobal Attention의 복잡도는 $O(l(r + l/k))$이다. [18]의 저자들은 실험적으로 $r = 127$, $k = 16$일 때 충분히 좋은 성능을 보인다고 주장한다.

LongT5 모델은 기존 Transformer 구조의 모델에 비해 긴 문장 입력을 처리할 수 있어 대화 내의 이전 발화로부터 맥락을

[1] https://montrealcorpustools.github.io/Montreal-Forced-Aligner/

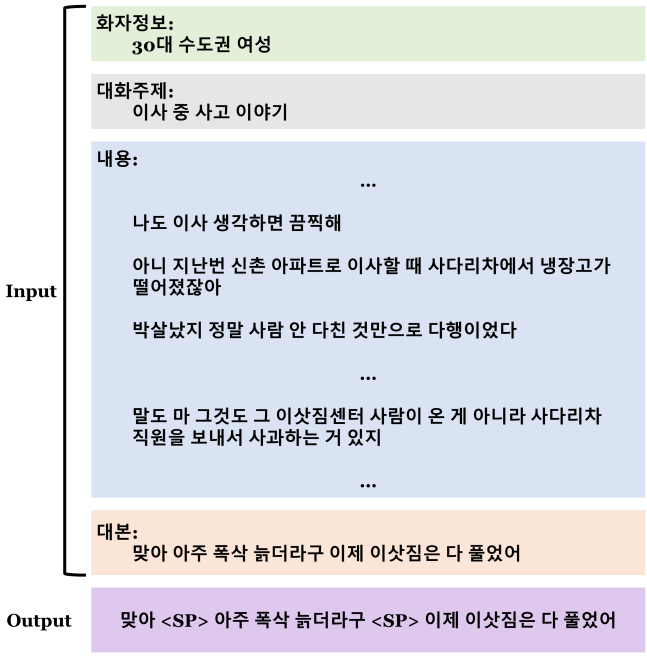

그림 2. 휴지 예측 모델의 데이터 입출력 예제

파악하여 휴지 위치를 예측하는 데 효과적이다. 이에 따라 본 연구에서는 영어 및 한국어 데이터셋으로 사전 학습된 Long-KE-T5[2]에 기반하여 대화 맥락에 기반한 휴지 예측 모델을 구성하였다.

모델은 그림 2와 같이 휴지 삽입 위치를 예측할 마지막 발화 데이터와 함께 화자 정보와 대화 주제 및 발화 내용을 입력으로 하여 휴지가 올바르게 삽입된 문장을 출력하도록 학습된다. 휴지 위치를 포함한 음성-음소 간 정렬 정보는 §2.1에서 소개한 Montreal Forced Aligner를 활용하여 라벨링하였다.

3.2 데이터

데이터셋	대화 데이터 수	발화 수 평균
Train	117622	19.3
Test	13144	20.0

표 1. 대화 데이터 통계

실험 및 평가를 위해 AI-Hub[3]에 공개되어 있는 '자유대화 음성(일반남여)'의 학습 및 평가 데이터셋을 활용하였다. 해당 데이터는 10대에서 50대 사이의 일반인 남녀 2,000여 명을 대상으로 수집한 4,000여 시간의 발화 음성 데이터로 구성되어 있으며, 동일한 주제 내에서 각 발화자에게 연속적으로 수집된

[2]https://github.com/AIRC-KETI/long-ke-t5
[3]https://www.aihub.or.kr/

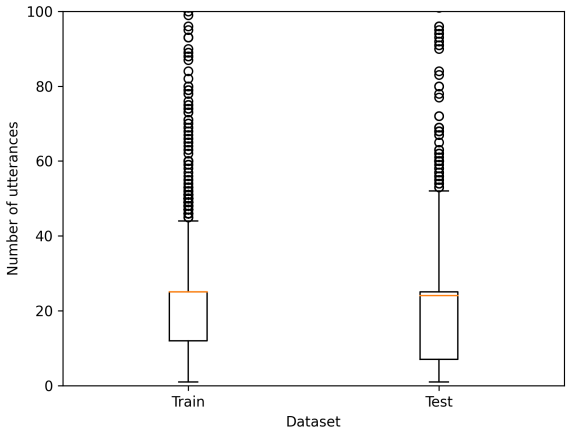

그림 3. 대화 데이터 내 발화 수 분포

음성을 취합하여 각각의 대화 데이터로 재구성하였다. 단, 발화 음성이 묵음이거나 파일 크기 10KB 미만 및 음성 길이가 1초 미만인 데이터는 이상치로 분류하여 제외하였다. 전처리 작업을 수행한 뒤 데이터 통계는 표 1 및 그림 3에서 확인할 수 있다.

4. 실험 및 결과

# of utterances	Precision	Recall	F1-score
0 (baseline)	78.23	35.12	48.48
1	78.65	35.38	48.80
5	78.19	35.51	48.84
10	78.88	36.34	49.76
20	77.68	**37.22**	**50.32**
50	**79.54**	35.79	49.36
100	79.22	35.93	49.44
300	79.52	35.22	48.82

표 2. 발화 수에 따른 휴지 예측 성능

표 2 및 그림 4는 이전 발화 수에 따른 휴지 예측 성능을 확인하기 위해 평가 데이터셋 상에서 평가를 진행한 결과이다. 이전 발화를 전혀 포함하지 않은 baseline에 비하여 모든 실험에서 F1-score가 향상되었으며, 20개의 이전 발화 데이터를 포함하여 휴지 예측을 수행한 모델의 경우 baseline 대비 F1-score 기준 최대 1.84%p 증가하였다. 특히 baseline 대비 최대 2.1%p 향상된 Recall 성능에 기인하여 모델이 휴지 삽입 위치를 재현하는데 주력하여 학습되었음을 확인하였다. 다만 발화 수가 증가할 수록 성능 향상 폭이 미미한 모습을 보이는데, 이는 발화 수가 많은 대화 데이터가 충분하지 않을뿐더러 대화가

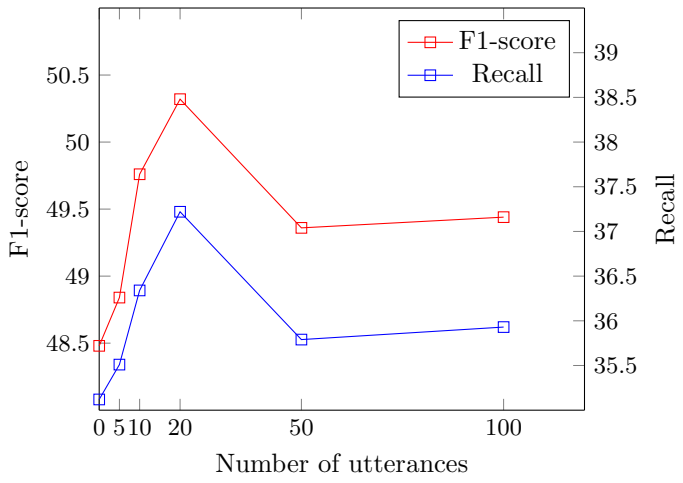

그림 4. 발화 수에 따른 휴지 예측 성능 그래프

길수록 다양한 세부 주제가 존재하므로 모델이 맥락을 파악하는데 어려움을 겪는 것으로 추정된다.

모든 실험은 NVIDIA RTX A6000 4기로 수행되었으며 각 모델은 32의 배치 크기로 약 15,000 step 동안 학습한 뒤 평가되었다. 또한 옵티마이저(Optimizer)는 AdamW, 학습률(Learning Rate)은 2e-5로 설정한 뒤 약 1,000 step 간의 warmup 후 선형적으로 감소하도록 조정하였다.

5. 결론

본 연구는 대화 맥락에 기반하여 적절한 위치에 휴지를 삽입하는 한국어 휴지 예측 모델을 제안하고 대화 데이터셋에서 검증한 결과를 보인다. 긴 대화 내에서 이전 발화를 최대한 포함하여 맥락 정보를 상세히 포착하고 휴지 예측에 활용할 수 있도록 Long-Input Transformer 기반의 LongT5 모델을 활용하였다. 실험을 통해 대화의 마지막 발화에서 휴지의 삽입 위치를 예측한 결과 기존 대비 F1-score 기준 최대 1.84%p 향상된 결과를 보였으며, 특히 Recall 성능이 강화됨에 따라 모델이 휴지 위치를 보다 정확하게 예측하도록 학습됨을 확인하였다.

향후 연구를 통해 자연언어처리 분야에서 비약적인 발전을 이룬 대규모 언어 모델에 적용하여 성능을 확인하고 지속 시간에 따라 세분된 휴지를 도입하여 자연스러운 음성 합성 시스템을 구축하고자 한다.

참고문헌

[1] OpenAI, "Gpt-4 technical report," 2023.

[2] B. Kim, H. Kim, S.-W. Lee, G. Lee, D. Kwak, D. H. Jeon, S. Park, S. Kim, S. Kim, D. Seo *et al.*, "What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers," *arXiv preprint arXiv:2109.04650*, 2021.

[3] C. McGinn and I. Torre, "Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots," *2019 14th ACM/IEEE international Conference on human-robot interaction (HRI)*, pp. 211–221, 2019.

[4] B. Sudharsan, S. P. Kumar, and R. Dhakshinamurthy, "Ai vision: Smart speaker design and implementation with object detection custom skill and advanced voice interaction capability," *2019 11th International Conference on Advanced Computing (ICoAC)*, pp. 97–102, 2019.

[5] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783, 2018.

[7] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[8] 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.

[9] G. Bailly and C. Gouvernayre, "Pauses and respiratory markers of the structure of book reading," *Interspeech 2012-13th Annual Conference of the International Speech Communication Association*, pp. Thu–O9d, 2012.

[10] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase break prediction for long-form reading tts: Exploiting

text structure information," 2018.

[11] V. Keri, S. C. Pammi, and K. Prahallad, "Pause prediction from lexical and syntax information," *Proceedings of International Conference on Natural Language Processing (ICON)*, 2007.

[12] Q. Shi, D. Jiang, F. Meng, and Y. Qin, "Combining length distribution model with decision tree in prosodic phrase prediction," *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[13] 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.

[14] 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.

[15] 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.

[16] J. Ainslie, S. Ontanon, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, and L. Yang, "Etc: Encoding long and structured inputs in transformers," *arXiv preprint arXiv:2004.08483*, 2020.

[17] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang *et al.*, "Big bird: Transformers for longer sequences," *Advances in neural information processing systems*, Vol. 33, pp. 17 283–17 297, 2020.

[18] M. Guo, J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang, "Longt5: Efficient text-to-text transformer for long sequences," *arXiv preprint arXiv:2112.07916*, 2021.

[19] J.-S. Hwang, S.-H. Lee, and S.-W. Lee, "Pausespeech: Natural speech synthesis via pre-trained language model and pause-based prosody modeling," *arXiv preprint arXiv:2306.07489*, 2023.