인공지능 기반 배리어프리 음성 화면 해설 제작 자동화 시스템 설계 및 구현

이혜준*, 박준형**, 윤태원*, 조재하***, 허나영****, 황대은*****, 유길상*

고려대학교*, 한국 뉴욕 주립 대학교**, 건국대학교***, 서울과학기술대학교****, 상명대학교*****

cometary01@korea.ac.kr, junhyeong.park@stonybrook.edu, yoontaewon1224@gmail.com, millaty24@gmail.com, hny030703@seoultech.ac.kr, softn121@gmail.com, ksyoo@korea.ac.kr

Design of an AI-Based Automated System for Producing Barrier-Free Audio Descriptions

Haejune Lee*, Junhyeong Park**, Taewon Yoon*, Jaeha ***, Nayeong Heo****, Daeeun Hwang*****, Gilsang Yoo*

Korea University*, Stony Brook University**, Konkuk University***, Seoul National University of Science and Technology****, Sangmyung University*****

요 약

시각장애인은 드라마와 영화 같은 시청각 콘텐츠를 완전히 이해하는 데 어려움이 있다. 이를 위해 배리어프리 동영상 콘텐츠 제작 자동화에 대한 연구가 필요한 실정이다. 기존의 음성 화면 해설 제작은 전문가의 작업에 의존하여 시간과 비용이 많이 소요되기에 콘텐츠 제작과 보급이 한정적이다. 이와 같은 한계를 보완하고자 본 연구는 인공지능 기술을 활용하여 음성 화면 해설 제작 과정을 자동화하고, 이를 통해 제작 효율성을 높이고 비용을 절감하며, 배리어프리 콘텐츠의 접근성을 확대하기 위한 시스템을 설계하고 구현하였다. 구현 결과, 제안한 시스템은 음성 화면 해설의 자동화를 통해 동영상 콘텐츠의 접근성을 개선하고, 모든 사람이 정보에 평등하게 접근할 수 있는 환경을 조성하는 데 이바지할 것으로 기대한다.

Ⅰ. 서 론

OTT(Over The Top)서비스의 급속한 발전과 함께 디지털 콘텐츠의 접근성에 대한 요구가 증가하고 있다. 특히 시각장애인을 비롯한 장애인들이 다양한 미디어 콘텐츠에 쉽게 접근하고 활용할 수 있도록 지원하는 기술의 개발이 중요한 과제로 떠오르고 있다. [1] 그 중에서도 동영상 콘텐츠의 시각적 요소를 음성으로 전달하는 '화면 해설'은 시각장애인들의 원활한 콘텐츠 감상을 위해 필수적이다. [2] 그러나 화면 해설의 제작 과정은 많은 시간과 비용이 소요되는 노동집약적 작업으로, 이에 대한 효율적인 해결 방안이 필요한 실정이다.

인공지능(AI: Artificial Intelligence) 기술의 발전은 이러한 문제를 해결할 수 있는 새로운 가능성을 열어주고 있다. [3] 특히, 자연어 처리(NLP: Natural Language Processing), 음성 합성(TTS: Text to Speech), 그리고 컴퓨터 비전(CV: Computer Vision)과 같은 기술들이 화면 해설의 자동화를 위한 핵심 요소로 주목받고 있다. [4] 이와 같은 AI 기술을 활용한 시스템을 구축하여 화면 해설의 제작 과정에서 투입되는 비용 및 노동력을 최소화하면서도, 빠르게 음성 화면 해설을 생성할 수 있다.

본 논문에서는 배리어프리 음성 화면 해설 제작의 자동화를 위해 시나리오와 자연어 처리 기술을 활용해 화면 해설용 텍스트를 생성하고, 비디오 콘텐츠의 이미지 정보를 처리하여 적절한 해설 구간을 도출하는 AI 기반 시스템을 제안한다. 이를 통해 기존의 화면 해설 제작 과정보다 시간과 비용의 측면에서 효율적인 방법론을 제시하고 구현하였다.

Ⅱ. 배리어프리 음성 화면 해설 자동화 시스템

제안한 배리어프리 음성 화면 해설 자동화 시스템의 흐름은 그림 1과 같다. 화면해설은 장소, 인물의 행동 등 시각적 정보를 해설해야 한다는 점에서 시나리오의 '지문'과 유사성을 가지기에 본 시스템은 영화 영상과 시나리오를 사용한다. 따라서 먼저 시나리오를 영화의 순서에 따라 재구성하고, 인물의 대사에 유의하여 음성 해설 타이밍을 추출한다. 다음으로, 패러프레이징 모델과 TTS모델을 활용하여 자동으로 음성해설을 생성하고 영상에 적용한다.

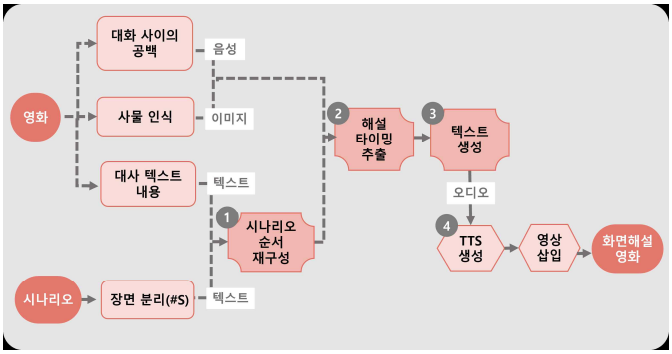


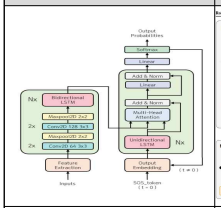
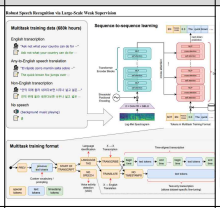
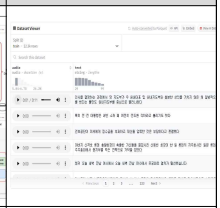
그림 1. 제안한 시스템 구성도

(1) 데이터 수집 및 전처리

먼저, 영화 대사를 텍스트로 변환한다. 각종 효과음과 음악이 섞인 영화의 소리에서 인물 간 대사를 추출하기 위해, 표 1에서와 같이 소음 환경에서도 좋은 전사(transcription) 성능을 보이는 STT(Speech to Text) 모델인 Whisper를 사용하였다.

대화가 이어지지 않는 구간에서 '아', '어'와 같은 단순한 텍스트가 반복적으로 출력되는 문제를 해결하기 위해 대화 감지(speech activity detection) 기술을 적용하고, 대화가 등장하는 구간과 그렇지 않은 구간을 분리하여 대화 구간만 Whisper의 입력으로 사용하였다.

표 1. 한국어 음성 처리 모델 비교 [5][6]

KoSpeech	Whisper	Whisper-medium-ko-zeroth
		
상대적으로 소음에 약함	소음에 강함 정제된 한국어 대화 + 일상 대화(약어, 속어 등 포함)로 학습	형식적이고 정제된 대화 위주인 Zeroth 데이터셋으로 Fine-tuning

(2) 시나리오 순서 재구성

화면 해설이 삽입될 적절한 구간을 찾기 위해 시나리오를 씬 별로 분리하고 대사와 지문을 구분하였다. STS(Sentence Textual Similarity) 태스크를 수행하는 SBERT모델을 활용해 시나리오의 대사 텍스트를 벡터로 변환 후, 영화에서 음성인식을 통해 추출된 대사 텍스트와의 문장 유사도를 계산하였다. 이를 통해 시나리오의 씬 순서를 영화와 일치하도록 정렬하고, 시나리오상에서만 존재하는 씬이 제거되도록 하였다.

(3) 해설 타이밍 추출

다음으로, 해설을 삽입할 정확한 타이밍을 결정하기 위해 씬 변경 지점을 알아내야 한다. 대사를 기준으로 씬의 변경 지점을 확인하되, 대사가 존재하지 않아 화면 해설 삽입 시점을 구분할 수 없는 장면 전환 지점은 컴퓨터 비전 기술을 사용하여 구분할 수 있도록 하였다.

(4) 텍스트 생성 및 Paraphrasing

위 과정을 통해 도출된 화면 해설 구간에 시나리오의 지문을 그대로 삽입할 수도 있으나, 공백 구간에 비해 지문이 길거나 완결되지 않은 문장으로 끝나거나 소리 정보에 대한 묘사를 포함하는 등 화면 해설용으로는 적합하지 않은 경우가 다수 존재하였다. 이를 해결하기 위해 BART모델(Bidirectional and Auto-Regressive Transformers)을 미세 조정하여 시나리오 지문이 화면 해설 형식에 맞게 패러프레이징 되도록 하였다. 미세 조정된 BART모델은 시나리오의 지문을 화면 해설 구간 길이를 기준으로 요약하고, 줌인, 줌아웃 등의 촬영 용어나 의성어를 제외하여 화면 해설용 텍스트를 생성한다.

그림 2. 원본 (좌)과 화면해설이 적용된 이후(우)의 프레임 및 음성, 영화 '분홍신' 中

(5) 화면 해설 텍스트 음성화

생성된 해설 텍스트를 TTS 기술을 활용하여 음성 파일로 변환한다. 이를 원본 영상에 삽입하여 화면 해설이 삽입된 영화를 완성한다.

(6) 구현 결과

구현된 시스템을 통해 자동 생성된 화면 해설이 적용된 결과는 그림 2와 같다. 대사가 부재한 장면에서는 시각장애인이 실시간으로 영화의 상황을 명확히 이해하기 어렵다. 그러나 그림 2에서 볼 수 있듯 대사가 없는 동안 화면 해설이 음성으로 제공되면 시각장애인의 콘텐츠 접근성을 높일 수 있다.

Ⅲ. 결론

본 연구에서는 기존의 수작업 방식과 비교하여 인력 투입을 배제하고도 배리어프리 영화 콘텐츠를 제작할 수 있는 AI 기반 화면 해설 자동화 시스템을 제안하였다. 제시된 방법론을 활용하여 다양한 배리어프리 영화 콘텐츠를 단기간에 제작하는 데 활용할 수 있다.

본 연구를 통한 AI기반 음성 화면 해설 제작 자동화는 배리어프리 영화 보급을 촉진하고, 시각장애인의 영화 감상 경험을 개선하는 데 기여할 수 있을 것으로 기대된다. 향후 연구에서는 AI 기술의 고도화뿐만 아니라, 다양한 장르에 걸친 해설 자동화 적용 방안을 탐구하고, 실제 사용자들의 피드백을 반영한 후속 연구가 진행될 예정이다.

ACKNOWLEDGMENT

본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 대학·기업 협력형 SW아카데미 사업의 지원을 받아 고려대학교 지능정보 SW아카데미 프로젝트 결과물로 수행되었습니다.

참 고 문 헌

[1] J. H. Choi, C. H. Ahn, J. Seo and O. Kwon, "Sensory effect representation for barrier-free broadcasting service," 2017 19th International Conference on Advanced Communication Technology (ICACT), PyeongChang, Korea (South), 2017, pp. 664-667, doi: 10.23919/ICACT.2017.7890176.

[2] 나준기. (2013). 화면해설방송과 배리어프리영화의 연출방법연구- 부산국제영화제 배리어프리영화 제작 중심으로-. 예술과 미디어, 12(4), 253-269.

[3] 유길상, 김현철. (2023-11-23). 배리어 프리 동영상 콘텐츠 자막생성을 위한 딥러닝 기반 배경음악의 감성 분류 모델 연구. Proceedings of KIIT Conference, 제주.

[4] L. Chen, Y. Deng, X. Wang, F. K. Soong and L. He, "Speech Bert Embedding for Improving Prosody in Neural TTS," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6563-6567, doi: 10.1109/ICASSP39728.2021.9413864.

[5] Soohwan Kim, Seyoung Bae, Cheolhwang Won, "Open-source toolkit for end-to-end Korean speech recognition, Software Impacts," Vol 7, 2021. https://doi.org/10.1016/j.simpa.2021.100054.

[6] Alec Radford, J. W. Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision". Proceedings of the 40th Int. Conf. on Machine Learning, PMLR 202:28492-28518,2023. https://doi.org/10.48550/arXiv.2212.04356.