chatGPT를 활용한 AI음성 수화 번역 서비스

김가희[1], 김지현[1], 김채민[1], 김민재[2], 박명수[3]
[1]성신여자대학교 AI융합학부 학부생
[2]서울과학기술대학교 컴퓨터공학과 학부생
[3]경기대학교 컴퓨터공학부 학부생

20211305@sungshin.ac.kr, 20211319@sungshin.ac.kr, 20211323@sungshin.ac.kr,
park45er@seoultech.ac.kr, watermelon@kyonggi.ac.kr

AI voice sign language translation service using chatGPT

Ga-Hee Kim[1], Ji-Hyeon Kim[1], Chae-Min Kim[1], Min-Jae Kim[2], Myeong-Soo Park[3]
[1]Dept. of AI Convergence, Sung-shin Women's University
[2]Dept. of Computer Engineering, Seoul National University of Science and Technology
[3]Dept. of Computer Engineering, Kyong-Gi University

요 약

본 연구는 농인의 언어권 보장을 위한 한국어-한국수어 번역 프로그램의 필요성을 제기하고 있다. 이를 위해 KoBART 모델을 활용하여 한국어 텍스트를 한국수어로 효과적으로 변환한다. ControlNet을 통해 수어 영상에서 손의 위치와 제스처를 정밀하게 추출하여 Stable Diffusion 모델을 제공함으로써 고해상도의 아바타 영상을 생성한다. 이러한 기술을 바탕으로 개발된 애플리케이션은 사용자가 음성을 입력하면 이를 텍스트로 변환하고, 변환된 텍스트에 대응하는 수어 영상을 순차적으로 재생하여 농인의 의사소통을 보다 원활하게 지원한다.

1. 서론

Covid -19 상황에서 정부의 통역사 배치가 늘어나는 사례로 보아 최근 몇 년간 농인의 언어권을 보장하기 위한 사회적 변화가 이루어지고 있다. 하지만 공공영역에서의 수어 통역 지원은 초기 단계에 머물러 있어 농인의 일상생활에서의 의사소통 편의성이 저해되고 있다.

국립국어원의 한국수어 활용 조사 통계에 따르면, 2022년 5월 기준, 농인의 수(추정 52,107명) 대비 수어 통역사 수(1,948명)는 약 3.7%에 불과하며, 수어 통역사 1명당 평균적으로 농인 26.7명을 담당하는 것으로 나타났다. 또한 수어통역 서비스를 받은 월평균 횟수가 약 50%가 3회 이하로 낮은 수치를 확인할 수 있다.

이에 따라 한국어-한국수어 번역 프로그램의 필요성이 대두되고 있다. 현재 상용화된 시스템들은 주로 수어를 한국어로 번역해주는 시스템이 많으며 음성 한국어를 이해할 수 있도록 돕는 서비스는 여전히 부족한 상태이다. 따라서 본 연구는 음성을 실시간으로 수어로 번역하는 딥러닝 기반의 서비스를 개발하여 농인의 의사소통을 지원하고 일상생활에서의 편의성 증진에 기여하고자 한다.

2. 적용 기술

2.1. 한국어-한국수어 변환 알고리즘

2.1.1 KoBART

KoBART는 BART 기반의 한국어 텍스트 처리 모델로, 인코더-디코더 구조를 갖춰 문맥 이해에 탁월하며 노이즈 제거를 통해 번역 및 요약과 같은 작업 성능을 최적화[1]한다. 또한, Text Infilling 방식을 적용하고 40GB 이상의 한국어 텍스트에 대해 학습되어 기존의 KoBERT보다 번역, 요약, 감성분석 등의 성능[2]이 우수하다.

본 연구에서는 국립국어원의 한국어-한국수어 말뭉치 데이터를 사용하였다. ChatGPT를 활용한 텍스트 데이터 증강을 통해 7,108개의 원본 데이터를 35,540개로 확장하여, KoBART translation model을 fine-tuning하였다. 이를 통해 한국어 텍스트를 입력받은 후 일치하는 한국 수어를 추출하는데 활용하였다. 학습 과정에서는 한국어 텍스트와 수어 텍스트를 일관된 입력 형식으로 토큰화하여 모델이 효율적으로 학습할 수 있도록 적용하였다. 모델의 핵심인 Seq2Seq 학습 파라미터와 Text Infilling 기술을 적용하여 문장의 누락된 토큰을 예측하고 복잡한 문장 구조를 처리함으로써 한국어 텍스트와 수어를 효과적으로 번역할 수 있었다.

2.2 포즈 추출 (Pose Extraction)

2.2.1 ControlNet

ControlNet은 기본적으로 대규모 사전 학습된 텍스트-이미지 확산 모델에 공간 컨디셔닝 제어를 추가하여, 더욱 정교한 이미지 생성을 가능하게 하는 end-to-end 신경망 아키텍처이다.[3]

Stable Diffusion의 U-Net 구조에서 인코더와 중간 블록을 고정(freeze)한 후 학습 가능한 복사본을 추가하여 사전 학습된 모델의 파라미터를 보존하면서 새로운 조건 제어를 학습한다. zero convolution을 사용하여 학습 초기 단계에 발생하는 해로운 noise를 제거함으로서 이미지 생성 과정을 최적화한다.

수어 영상에서 손의 위치와 신체의 제스처를 정밀하게 추출하고 해석하기 위해 ControlNet의 공간적 제어가 필수적이며, 이는 다양한 조건부 입력을 통해 구현된다. 본 연구에서는 control-openpose로 포즈와 관절 키포인트의 위치를 정의하고, control-softedge는 영상 내 손 객체의 외곽선과 윤곽선을 정의함으로써, 영상 생성 시보다 정교한 구조적 정보를 추출하고 아바타의 생성 위치를 고정하였다.

2.3 영상 생성 (Video Generation)

2.3.1 Stable Diffusion

아바타 영상 생성을 위해 사용된 모델은 대표적인 text-to-image model인 Stable Diffusion model이다. Stable Diffusion은 프롬프트를 기반으로 고해상도의 이미지를 생성하는 것으로 잘 알려져있는 모델이다.[4] 이를 영상 제작에 사용되기 위하여 frame 단위로 영상을 이미지로 분할하여 diffusion model에 적용하였다.

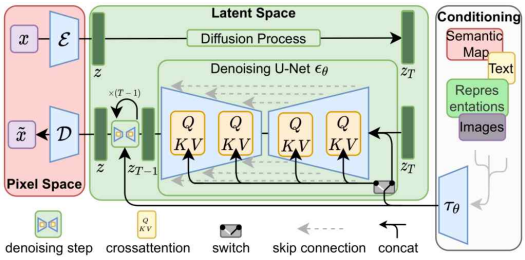

그림 1 Stable Diffusion 구조

Stable Diffusion의 기본이 되는 check-point model로는 majicMIX_realistic_v6 model을 활용하여 현실감을 나타내는 아바타를 생성하는 데 집중하였다. VAE 모델은 Stability.ai 사의 표준 VAE인 FT-MSE-8400000을 활용하여 각 프레임 당 이미지 생성 품질을 향상시키는데 활용하였다. 이후 LoRA 모델을 추가하여 디테일을 높이는 데 활용하였다.

해당 모델의 ComfyUI workflow와 적용한 결과는 아래와 같다.

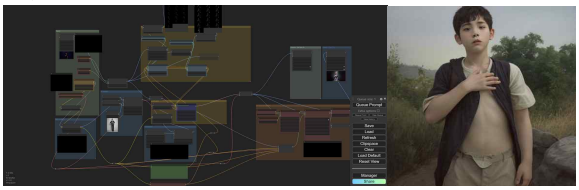

그림 2 ConfyUI workflow와 생성 결과

2.3.2 Stable Diffusion과 ControlNet의 연관성

Stable Diffusion은 프롬프트를 기반으로 이미지를 생성하는 강력한 모델이지만, 구조·공간적인 정보를 반영하는 데 한계가 있다. 따라서 ControlNet을 통하여 추출된 포즈 정보를 Stable Diffusion에 제공하여 이미지 생성 과정에서 기존의 수화 동작을 유지하도록 한다. 아래는 ControlNet과 Stable Diffusion을 적용한 예시이다.

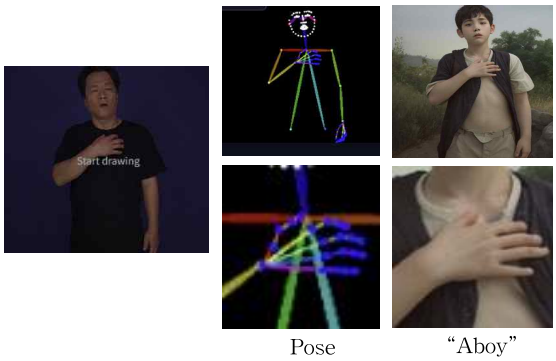

표 2 ControlNet, Stable Diffusion 적용 결과

3. 시스템 구현

본 애플리케이션은 농인의 음성으로부터 텍스트를 추출하고, 이를 바탕으로 수어 영상을 재생할 수 있도록 설계되었다. 사용자로부터 음성을 입력받아 이를 한글 텍스트로 변환한 후, 변환된 텍스트에 대응하는 수어 아바타 영상을 재생한다. 음성 인식에는 Speech-to-Text 플러그인을 사용하며, 이를 통해 실시간으로 음성을 텍스트로 변환한다. 이후, 변환된 텍스트를 분석하여 이에 해당하는 수어 아바타 영상을 생성하여 재생하는 구조이다.

생성된 수어 영상 데이터를 Google Firebase에서 불러올 수 있으며, 사용자가 입력한 음성 및 이에 대응하는 텍스트와 수어 영상 데이터를 Firebase를 통해 관리하고 확인할 수 있다. 이 애플리케이션은 농인의 의사소통을 보다 원활하게 지원하기 위해 개발되었으며, 직관적인 사용자 인터페이스를 제공함으로써 사용자가 쉽게 접근할 수 있도록 하였다.

4. 결론

공공영역에서의 수어 통역 지원이 여전히 초기 단계에 머물고 있어 농인의 일상생활에서 의사소통의 편의성이 저해되고 있는 상황이다.

이러한 문제를 해결하기 위해 한국어-한국수어 번역 프로그램의 필요성이 대두되었으며, 본 연구에서는 딥러닝 기반의 실시간 음성-수어 번역 서비스를 개발하였다. 국립국어원의 한국어-한국수어 말뭉치를 학습데이터로 사용하였으며, chatGPT를 활용한 데이터 증강을 통하여 부족한 데이터의 양을 보완하였다. KoBART 모델을 미세튜닝하여 한국어 텍스트를 수어로 효과적으로 번역할 수 있는 시스템을 구축하였고, ControlNet과 Stable Diffusion을 통해 정교한 수어 영상을 생성하는 방법을 제안하였다. 이 과정에서 손의 위치와 신체의 제스처를 정확하게 추출하고, 이를 기반으로 고해상도의 수어 영상을 생성함으로써 시각적으로도 농인 사용자에게 친숙한 형태의 의사소통 방법을 제공할 수 있었다. 본 애플리케이션은 농인이 음성을 텍스트로 변환하고, 그에 해당하는 수어 영상을 순차적으로 재생할 수 있도록 설계되었다.

결론적으로, 본 연구는 농인의 의사소통 편의성을 증진시키기 위한 기술적 기초를 마련하였으며, 이를 통해 농인들이 다양한 매체를 통해 더 쉽게 정보를 얻을 수 있는 환경을 조성하는 데 기여한다.

본 연구는 향후 감정 및 표정의 수어 비수지 표현을 추가하여, 보다 정확도 높은 서비스 구현을 목표로 한다. 또한 의료, 보건, 엔터테인먼트 등 다양한 분야로 확장될 수 있으며, 궁극적으로 농인의 정보 접근성을 향상시키고 그들의 권리와 기회를 증진시키는 데 기여할 것이다.

참고문헌

[1] 구다훈, "생성요약의 사실 불일치 문제 개선을 위한 관련성과 중복성을 고려한 손실 함수 기반의 KoBART 모델", 한국정보기술학회논문지, 제20권, 제12호, 2022년

[2] 이민아, "KoBERT, KoGPT-2, KoBART 활용 및 하이퍼파라미터 최적화를 진행한 리뷰 감성분석 애플리케이션 구현", 디지털콘텐츠학회논문지, 제24권, 제11호, 2023년도

[3] Lvmin Zhang, Anyi Rao, Maneesh Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," arXiv preprint arXiv:2302.05543, 2023.

[4] Rombach et al., High-Resolution Image Synthesis With Latent Diffusion Models, CVPR 2022

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.