영상 음성 기반 한국어-글로스 변환과 글로스-포즈 매핑을 이용한 한국 수어 아바타 번역 시스템

황시연, 엄성용
서울여자대학교

movoa3@swu.ac.kr, osy@swu.ac.kr

A Korean Sign Language Avatar Translation System Using Speech-to-Text, Text-to-Gloss, and Gloss-to-Pose Mapping

Hwang Si Yeon, Ohm Seong Yong
Seoul Women's University

요 약

본 논문은 영상의 음성 정보를 입력으로 받아 한국어 전사(STT), 글로스 변환(Text-to-Gloss), 글로스-포즈 매핑(Gloss-to-Pose)에서 3D 아바타 애니메이션 생성으로 이어지는 일련의 프로시저(Procedure)를 이용하여 한국수어(KSL) 번역 시스템을 구현한다. 본 시스템의 구성은 (1) 영상 업로드 및 오디오 추출, (2) STT를 이용한 한국어 문장 생성, (3) T5 기반 한국어-글로스 시퀀스 생성, (4) 글로스-포즈 인덱스를 이용한 포즈 벡터 시퀀스 구성, (5) Unity 기반 아바타 렌더링 등의 5개의 프로시저(Procedure)로 구성된 모듈형 파이프라인 형태이다. 문장-글로스 성능은 총 12,142개 샘플의 평가셋에서 BLEU 0.3266, ROUGE-L 0.8027, 토큰 단위 F1 0.6696을 기록하였다. 그리고 평균 예측 길이(14.89 토큰)는 평균 정답 길이(16.90 토큰)보다 짧아, 일부 표현 누락 및 동의 표현 변이의 영향 가능성을 확인하였다. 본 논문에서는 자막이 없는 영상의 음성 정보를 한국수어 아바타 동작으로 변환하는 파이프라인을 제안하며, 이를 통해 농인 사용자가 자막에 의존하지 않고 영상을 수어로 직접 이해할 수 있는 접근성 향상을 지향한다.

Ⅰ. 서 론

일반적인 자막이 없는 영상도 농인이 수어로 바로 이해할 수 있도록 영상의 음성 정보를 한국수어 표현으로 변환해 수어 화면을 자동으로 생성하여 제공하는 시스템을 구현하였다. 농인은 동영상을 시청할 때, 자막이 제공되지 않으면 내용의 핵심을 파악하기 어렵고, 자막이 제공되더라도 언어적 친숙성의 한계로 인해 이해도가 충분하지 않을 수 있다. 따라서 본 연구는 사용자 영상 입력을 받아서 음성을 추출한 뒤, STT를 통해 한국어 텍스트를 생성하고, 이를 T5 기반 Text2Gloss 모델로 글로스 시퀀스로 변환한다. 이후 글로스-포즈 인덱스를 통해 포즈 벡터를 구성하여 Unity 아바타로 수어 동작을 재생하는 STT-Text2Gloss-Gloss2Pose-Avatar 파이프라인을 제안한다. 특히 문장에서 글로스를 변환하는 성능이 전체 시스템 품질을 좌우함으로, 본 논문에서는 데이터 구축 및 정규화 방법을 포함해 Text2Gloss 모듈의 성능을 평가한다.

Ⅱ. 본론

2.1 시스템 구성

본 시스템은 영상에서 오디오를 추출한 뒤, STT로 한국어 문장을 생성하고, T5 기반 Text2Gloss 모델로 글로스 시퀀스를 생성한다. 이후 글로스 토큰을 글로스-포즈 인덱스에서 조회하여 포즈 벡터 시퀀스를 구성하고 Unity에서 아바타 애니메이션으로 재생한다. 이 구조는 모듈 단위 교체가 가능하여, 향후 STT 엔진 교체 또는 Gloss2Pose의 생성형 모델 확장에 유리하다. 그림 1은 음성에서 글로스 시퀀스를 생성하는 전반부 변환 과정을, 그림 2는 생성된 글로스를 동작 벡터 및 아바타 애니메이션으로 매핑하는 후반부 파이프라인을 개략적으로 나타낸다.

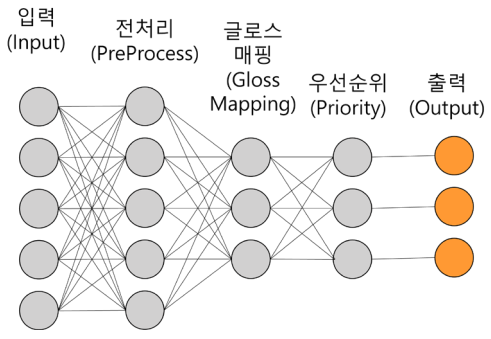

그림 1. Text2Gloss 파이프라인

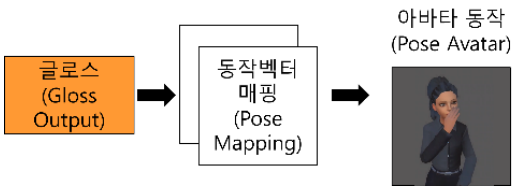

그림 2. Gloss2PoseAvatar 파이프라인

2.2 글로스(Gloss)의 정의 및 역할

글로스는 수어 동작을 학습 및 분석하기 위해, 수어의 의미 단위를 텍스트 토큰으로 표기한 중간 표현이다. 수어 번역/생성에서는 한국어 문장과 수어 표현 간 어순 및 문법 구조가 크게 차이점을 보이기에, 한국어 문장을 글로스 시퀀스로 1차 변환한 뒤 글로스에 대응되는 동작(포즈)을 생성하는 방식을 사용한다. 본 시스템 또한 '한국어 문장, 글로스, 포즈 시퀀스'의 흐름으로 변환을 수행한다. 표 1은 한국어 문장이 수어 글로스 시퀀스로 변환되면서 어순과 표현 방식이 어떻게 달라지는지를 예시로 보여준다.

표 1. 한국어 문장-글로스 변환 예시

기존 문장	만약 1종 면허증을 땄는데도 불구하고 버스 운전이나 택시 일을 농인이 안 하면요?
글로스 변환 문장	만약1 일종1 자격증1 가지다2 끝1 불구하다1 버스1 운전2 택시1 농1 불가능1

2.3 Text2Gloss(문장-글로스) 변환 모델

영상에서 추출된 한국어 문장을 입력으로 하여, Transformer 계열의 T5-base 모델을 미세조정을 함으로써 한국어 문장에서 한국수어 글로스로 시퀀스 변환을 수행하였다. 데이터는 AI-Hub 및 국립국어원 자료에서 문장-글로스 쌍을 추출, 정규화하여 약 14만 개 규모로 구축하였고, 이를 훈련/검증 9:1 비율로 나누어 학습하였다. 모델 출력은 공백으로 구분된 글로스 토큰 시퀀스이며, 전처리 단계에서 특수기호 제거, 문장 분리 및 정규화 등을 적용하였다.

2.4 Gloss2Pose 매핑 및 아바타 구동

생성된 글로스 토큰은 사전에 구축한 글로스-포즈 인덱스(JSON)에서 조회하여 포즈 벡터 시퀀스로 변환된다. 인덱스는 수어 영상으로부터 MediaPipe 기반으로 표정, 상반신, 양손 좌표를 프레임 단위로 추출한 뒤, 글로스 구간에 해당하는 포즈 시퀀스를 정규화하여 글로스 단위로 저장함으로써 구축하였다. 런타임에서는 각 토큰에 대응하는 포즈 시퀀스를 시간 순서대로 결합하여 문장 단위 포즈 시퀀스를 구성하며, 최종 포즈 시퀀스는 Unity 기반 리깅 모델에 적용되어 3D 아바타 애니메이션으로 재생된다. 본 연구에서 구현상 중점은 Text2Gloss 모델 학습과 글로스 포즈 인덱스 설계에 있다.

2.5 문장-글로스 평가 및 결과

문장-글로스 모듈은 정답을 알고 있는 총 12,142개 샘플의 평가셋에서 BLEU, ROUGE-L, 토큰 단위(Precision, Recall, F1) 등으로 성능을 측정하였다. 표 2는 해당 평가셋에서 제안한 문장-글로스 번역 모델이 지표별로 어느 정도의 성능을 보였는지를 요약한 것이다.

표 2. 문장-글로스 성능 평가 결과

지표	값
BLEU	0.3266
ROUGE-L	0.8027
Token Precision	0.7148
Token Recall	0.6298
Token F1	0.6696
Avg. Pred Length	14.89
Avg. Ref Length	16.90

BLEU는 0.3266으로 N-gram 기반 관점에서 중간 수준의 일치도를 보였다. 다소 낮은 값이지만, 수어 글로스 번역에서 단어의 순서 변동이 크다는 점을 고려한다면 의미 있는 수치라고 할 수 있다. ROUGE-L은 0.8027로 상대적으로 높게 나타나, 평가셋 기반으로 모델이 예측한 글로스가 정답 글로스와 상당 부분 일치함을 보여준다. 한편, 평균 예측 길이(14.89 토큰)가 평균 정답 길이(16.90 토큰)보다 짧아 일부 표현 누락 가능성이 존재하며, 이는 짧은 문장보다는 단어 수가 많은 긴 문장에서 상대적으로 성능이 떨어지는 경향과도 연결된다. 전반적으로 모델은 핵심 글로스는 비교적 잘 예측하지만, 부가적인 글로스를 생략하거나 더 짧은 표현으로 대체하는 경향을 보인다.

Ⅲ. 결론

본 논문에서는 영상에서 추출한 음성 정보를 바탕으로 한국수어 아바타 동작을 생성하는 STT-Text2Gloss-Gloss2Pose 파이프라인을 설계하고, 이를 구현하였다. 총 12,142개의 검증 데이터를 대상으로 평가한 결과 BLEU 0.3266, ROUGE-L 0.8027, 토큰 F1 0.6696을 확인하였다. 이는 학습 데이터가 제한적인 환경에서도 글로스 토큰 수준에서 비교적 안정적인 예측이 가능함을 보여준다. 앞으로는 글로스 표준화와 데이터 규모 확장을 진행하고, 표정이나 입 모양과 같은 비수지 요소를 포함하여 번역 품질과 동작의 자연스러움을 개선할 계획이다.

ACKNOWLEDGEMENT

본 연구는 서울여자대학교 SW 중심대학추진사업단의 지원의 연구결과로 수행되었음 (2025)

참 고 문 헌

[1] 서울특별시사회복지협회, "수어통역사", 2022.07.12, https://sasw.or.kr/interview/613643.

[2] 이기영, 신종윤, 임수중, 권오욱, "한국어 토큰-프리 사전학습 언어모델 KeByT5를 이용한 한국어 생성 기반 대화 상태 추적", 제35회 한글 및 한국어 정보처리 학술대회 논문집, pp. 644-647, 2023. 10.

[3] 최지훈, 이한규, 안충현, "아바타 수어 서비스를 위한 한국어-한국수어 변환 기술 연구", 한국방송미디어공학회 2020 하계학술대회 학술발표대회 논문집, pp. 337-338, 2020.07.