Term Project_가상 인간의 통역 및 제스처 생성 시스템 구현

120220354 윤수인

1. Introduction

최근 각광받고 있는 메타버스 분야에서 여러 대기업들이 AI 기술을 적용해 메타버스 속 다양한 공간에서 활약할 수 있는 버추얼 휴먼에 투자하는 사례가 늘고 있다. 현재의 가상 인간은 주로 기업 위주의 서비스, 광고에 사용되고 있으나 메타버스의 진흥을 위해서는 개인화가 필요하다. 이 연구는 가상 인간이 실시간으로 통역을 하고 제스처를 생성하는 것을 통해 사용자가 메타버스 내에서 언어에 국한되지 않고 다양한 언어권의 사람들과 쉽게 소통할 수 있게 함으로써 메타버스 산업의 발전에 기여할 수 있다.

이 시스템을 구현하기 위해 Unity 에 3D 아바타를 구현하고 OpenAI 의 Whisper 를 사용하여 유저의 마이크 인풋으로 들어온 음성을 받아 번역하고 Text 로 받아와 (STT) Gesticulator 를 사용하여 제스처를 생성하고 Text 를 다시 TTS를 사용하여 음성으로 송출하는 방식을 사용했다.

2. Related Works

2.1. OpenAI Whisper

Whisper는 웹에서 수집된 680,000시간의 다국어 및 멀티태스킹 supervised 데이터에 대해 훈련된 자동 음성 인식(ASR) 시스템이다.

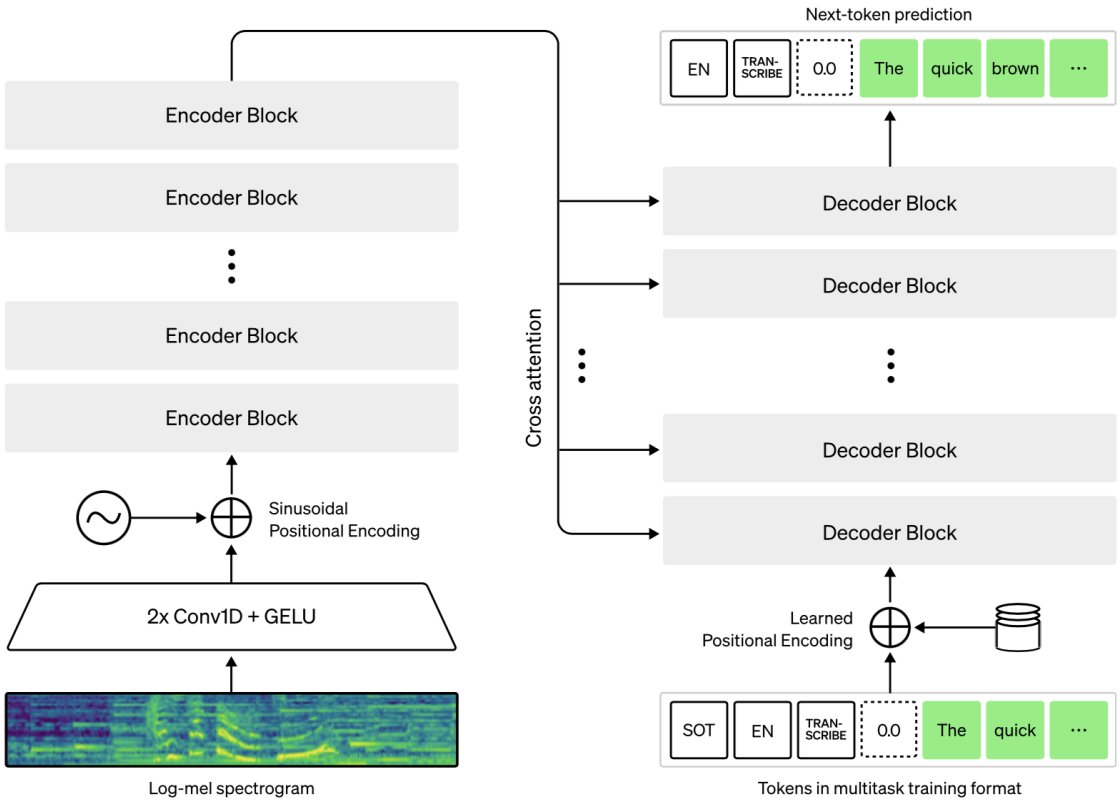


Figure 1. Whisper's Architecture

위스퍼의 구조는 encoder-decoder Transformer로 구현된 간단한 end-to-end 접근 방식이다. 입력 오디오는 30초 청크로 분할되어 log-Mel spectrogram으로 변환된 다음 인코더로 전달된다. decoder 는 단일 모델이 언어 식별, 구문 수준 타임스탬프, 다국어 음성 전사 및 영어 음성 번역과 같은 작업을 수행하도록 지시하는 특수 토큰과 혼합되어 해당 텍스트 캡션을 예측하도록 훈련된다.

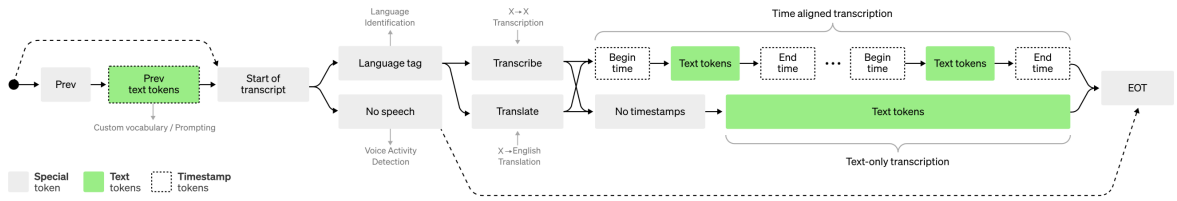


Figure 2. Whisper's Flow

2.2. Gesticulator

Gesticulator는 손짓과 표정과 같은 비언어적 행동 생성을 위한 기계 학습 모델로 오디오와 텍스트 데이터를 인풋으로 사용하여 적절한 제스처를 결합 각도 회전의 시퀀스로 생성할 수

있다. 가상 에이전트와 휴머노이드 로봇 모두에 적용될 수 있다.

2.3. Google-Cloud-TextToSpeech(gTTS)

DeepMind의 음성 합성 전문 기술을 기반으로 제작되어 API가 인간과 흡사한 수준의 음성을 제공하며 40개 이상의 언어 및 방언을 지원하는 220여 개의 음성 중 선택할 수 있다.

2.4. Pythonnet

Python.NET은 C# 기반의 unity 와 파이썬의 원활한 통합을 제공하는 패키지이다. 파이썬 코드가 CLR과 상호 작용할 수 있게 하며, 파이썬을 삽입하는 데에도 사용될 수 있다.

2.5. BVH Retargeting

Gesticulator에서 생성된 BVH 파일을 가상 인간에 retargeting하여 적절한 제스처를 취하도록 한다.

3. Materials and Methods

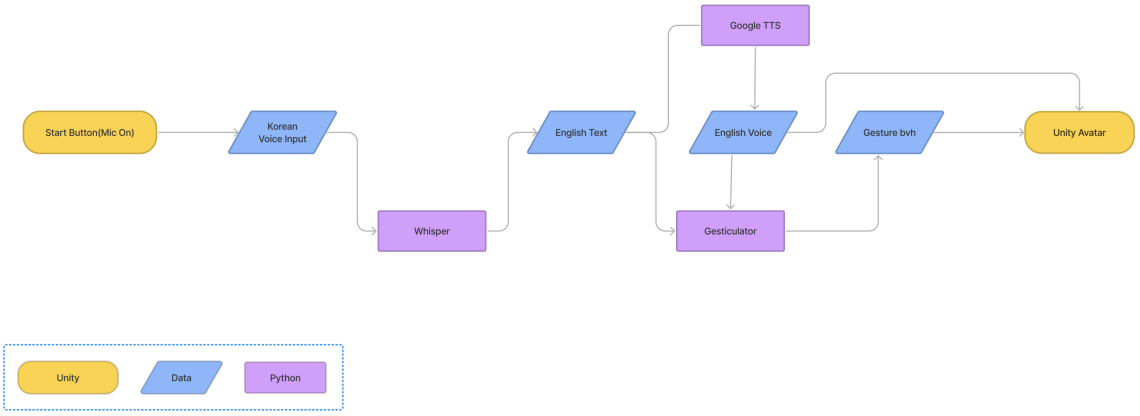


Figure 3. Flow Chart

OpenAI 의 Whisper 를 사용하여 유저의 마이크 인풋으로 들어온 음성을 받아 번역하고 Text 로 받아와 (STT) Gesticulator 를 사용하여 제스처를 생성하고 Text 를 다시 TTS를 사용하여 음성으로 송출한다.

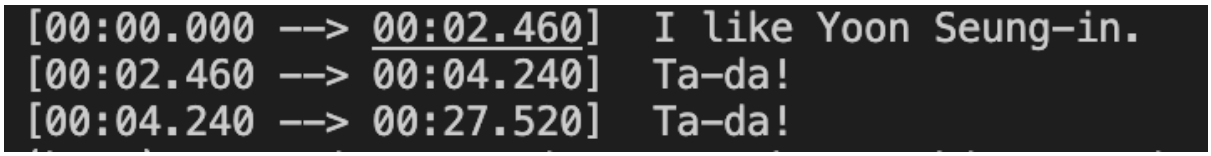
1. Unity 마이크를 통해 한국어 음성을 녹음해 wav파일 생성.

2. Python Whisper를 이용해 1에서 얻은 한국어 음성을 영어로 번역 후 txt 파일 생성.
3. Python gTTS를 통해 2의 txt 파일에서 영어 음성 wav 파일 생성.
4. Python 영어 txt 파일과 wav파일이 Gesticulator의 인풋으로 들어가 bvh 파일 생성.
5. Unity bvh 파일을 unity avatar에 리타게팅하여 제스처를 동작하게 하고 wav를 재생.

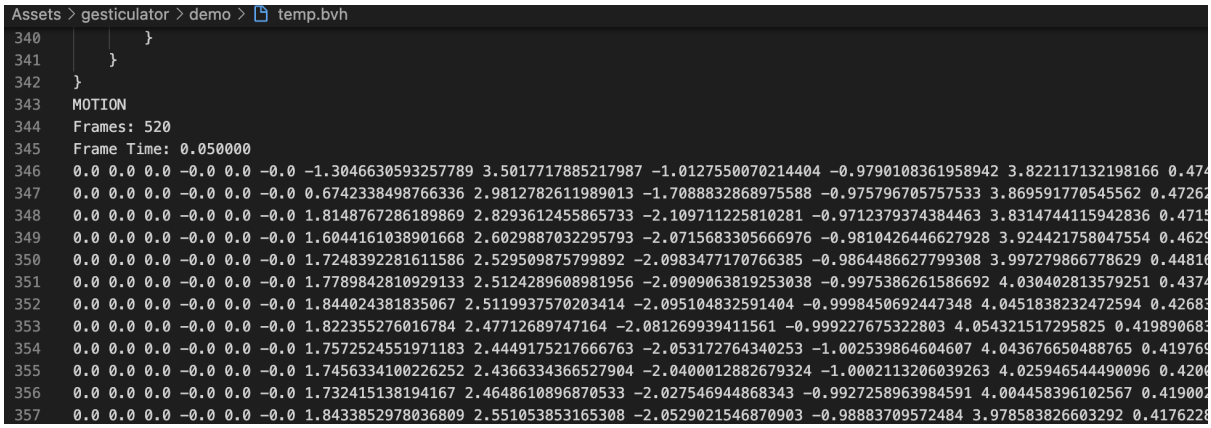
4. Results

1. wav 파일 생성

2. Whisper로 한국어 음성이 영어 text로 바뀌어 txt파일이 생성됨



4. temp.bvh 생성



5. Video Link: 최종 완성에 실패했다.

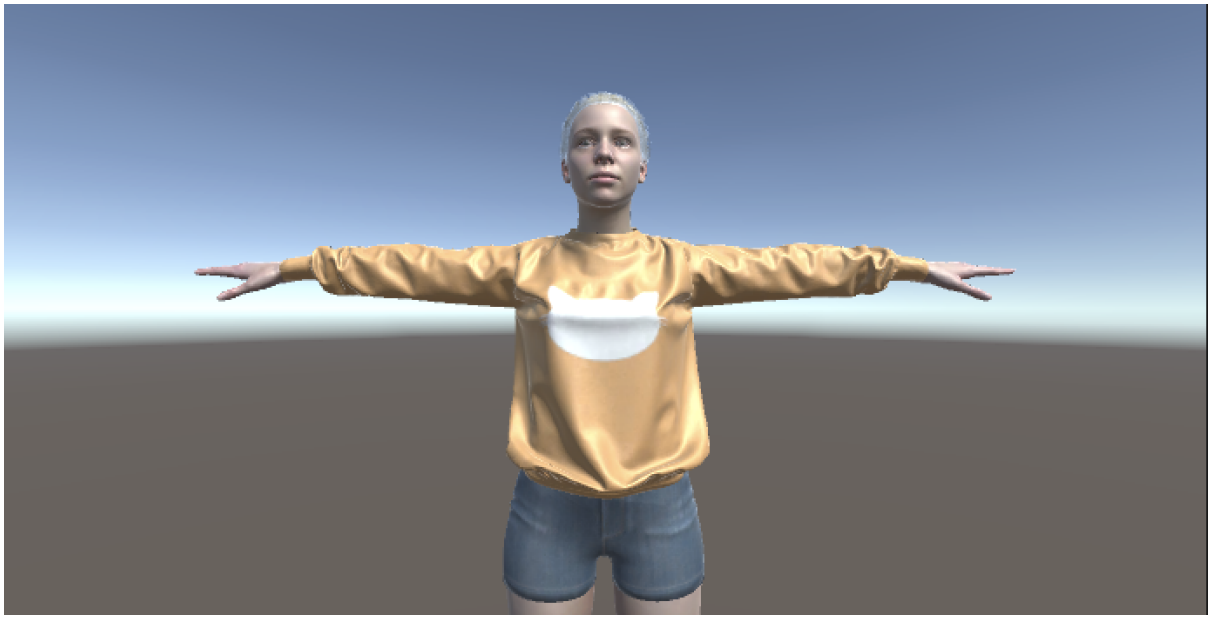


Figure 5. Avatar

5. Conclusion

본 연구를 통해 가상 현실에서 유저를 나타낼 수 있는 아바타가 자동으로 통역을 하고 그에 맞는 제스처를 취하는 것을 구현하고자 했으나 미흡한 실력 탓에 제대로 구현하지 못하였다.

나아가 표정을 생성해주는 모델과 더 매끄러운 통역, 제스처를 생성할 수 있는 모델을 사용한다면 메타버스에서 전 세계의 다양한 언어권의 유저들이 언어에 구애받지 않고 쉽게 커뮤니케이션하며 서비스를 즐길 수 있을 것으로 기대된다.