사회적 상호작용이 가능한 대화형 메타휴먼의 설계

김진학, 이영우, 최현범, 서정욱, 안현*
한신대학교 AI·SW학과
{wlsgkr99, duddn1124, chl9717, jwseo, *hyunahn}@hs.ac.kr

Design of a Socially Interactive and Conversational Metahuman System

Kim Jin Hak, Lee Youngwoo, Hyeon-Beom Choi, Jeongwook Seo, Hyun Ahn*
Dept. of Computing and AI, Hanshin Univ.

요 약

본 논문은 사용자의 음성과 얼굴 방향에 실시간으로 반응하는 대화형 메타휴먼 시스템을 설계하였다. 본 시스템은 음성 인식(ASR), 대규모 언어 모델(LLM), 텍스트-음성 변환(TTS), 얼굴 애니메이션 생성, 시선 추적 기술을 Unreal Engine 기반의 MetaHuman 캐릭터와 연동해 구성되었다. 사용자가 마이크로 질문하면, 언어 모델 기반 응답이 음성으로 합성되어 메타휴먼 캐릭터를 통해 발화되고, 동시에 웹캠을 통해 인식된 얼굴 방향 정보가 캐릭터의 고개 움직임으로 반영된다. 이와 같은 자동화된 구조는 몰입감 있는 상호작용 경험을 제공하며, 키오스크, 가상 상담, 교육 등 다양한 분야로의 확장이 가능하다. 본 연구는 이러한 시스템 설계의 실제 적용 가능성을 확인하고, 향후 감정 인식 및 다중 사용자 대응 기술을 통합하는 방향으로의 확장성을 제시한다.

Ⅰ. 서 론

최근 메타버스와 가상 공간 기술의 발전으로, 사용자의 언어, 표정, 시선에 실시간으로 반응할 수 있는 대화형 인터페이스 시스템에 대한 관심이 증가하고 있다. 이는 단순한 정보 전달을 넘어서, 현실 세계의 인간과 유사한 수준의 존재감과 상호작용 능력을 가진 디지털 아바타를 의미하며, 교육, 상담, 키오스크 안내, 가상 쇼핑, 전시 체험 등 다양한 실생활 분야에서 폭넓게 활용될 가능성을 가지고 있다.

기존의 대화형 시스템은 대체로 자동 음성 인식(Automatic Speech Recognition, ASR)과 텍스트 기반 응답, 그리고 단순한 음성 합성(Text-to-Speech, TTS)에 의존하는 구조를 지니고 있어 사용자와의 몰입감 있는 상호작용을 제공하는 데 한계가 있었다. 예를 들어, 응답하는 캐릭터가 사용자의 얼굴을 인식하지 못하거나 음성과 립싱크가 일치하지 않아 실제 사람과의 상호작용처럼 느껴지지 않는 문제가 존재하였다. 또한, 시스템 설계 상 실시간성 확보가 어렵고, 시선이나 감정과 같은 비언어적 신호에는 반응하지 못하는 기술적 제약도 존재한다.[1] 이러한 한계는 사용자로 하여금 캐릭터와의 상호작용이 기계적이고 단절되어 있다는 인상을 주며, 몰입을 방해하는 요인으로 작용한다.

본 연구의 목적은 이러한 기존 대화형 시스템의 한계점을 보완하여, 사용자와 실시간으로 자연스럽게 대화하고 시선을 맞추며 반응하는 몰입형 메타휴먼 인터페이스를 설계하는 데 있다. 이를 위해 음성 인식(ASR), 대규모 언어 모델(Large Language Model, LLM), 텍스트-음성 변환(TTS), 얼굴 애니메이션 생성 기술, 얼굴 방향 추적 기술 등을 통합하여, 음성과 비언어적 신호가 모두 반영되는 시스템 아키텍처를 설계하였다.[2]

논문에서는 먼저 전체 시스템 구조와 흐름을 설명한 뒤, 음성 인식, 대화 생성, 음성 합성, 얼굴 애니메이션, 시선 추적 등 세부 기술 요소들을 단계별로 정리하고, Unreal Engine 기반의 구현 방식을 통해 실시간 반응 구조가 어떻게 구성되는지를 살펴본다. 특히 각 기술의 연계 방식과 자동화된 데이터 흐름에 주목하며, 이들이 하나의 통합된 대화형 시스템에서 어떻게 유기적으로 작동하는지를 설명한다. 마지막으로 본 시스템의 적용 가능성과 향후 발전 방향을 제시한다.

Ⅱ. 대화형 메타휴먼 설계

1. 시스템 구조

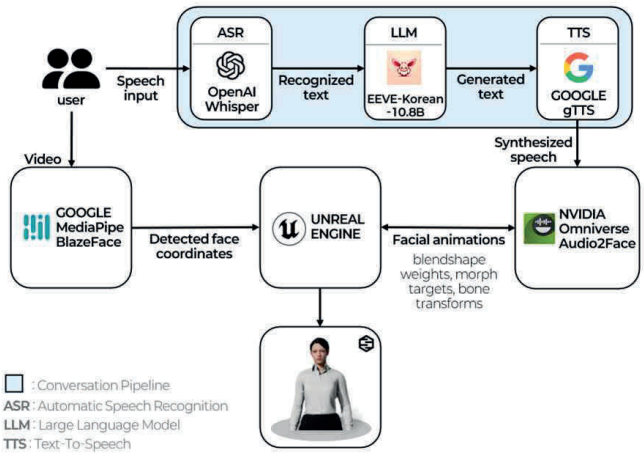

그림 1. 전체 시스템 구조

본 시스템은 크게 다섯 가지 모듈로 구성된다. Whisper를 이용한 음성 인식, EEVE-Korean-Instruct-10.8B 모델[3]을 이용한 자연어 응답 생성, gTTS를 통한 음성 합성, Audio2Face를 이용한 립싱크 기반 얼굴 애니메이션 생성, MediaPipe를 이용한 얼굴 방향 추적 및 Unreal Engine 기반의 메타휴먼 캐릭터 연동이다.

이 구조는 사용자가 마이크를 통해 음성 질의를 입력하면, Whisper가 이를 텍스트로 변환하고, EEVE-Korean-Instruct-10.8B 모델이 응답 문장을 생성한다. 해당 문장은 gTTS를 통해 음성으로 변환되며, 생성된 음성은 Audio2Face로 전달되어 립싱크 및 표정 애니메이션이 생성된다. 동시에 사용자의 얼굴 방향은 웹캠으로 촬영된 영상에서 MediaPipe가 인식하여 사용자의 얼굴 좌표값을 UDP 통신을 통해서 Unreal Engine으로 전달되며, 메타휴먼 캐릭터의 고개 움직임으로 반영된다.

2. 세부 단계별 적용 기술

2.1 음성인식(ASR): Whisper

OpenAI에서 개발한 Whisper는 다양한 언어의 음성을 텍스트로 변환할 수 있는 오픈소스 자동 음성 인식 모델이다. 본 연구에서는 이 모델을 사용하여 실시간으로 사용자 마이크 입력을 받아 텍스트로 변환하는 데 활용하였다.

2.2 대화생성(LLM): EEVE-Korean-Instruct 10.8B

이 모델은 야놀자에서 개발된 한국어에 특화된 대규모 언어 모델로, 오픈소스 기반으로 Hugging Face에서 제공되며 약 108억 개의 파라미터를 가진다. Whisper로부터 전달된 텍스트를 기반으로 자연스러운 한국어 응답을 생성한다.

2.3 음성합성(TTS): gTTS

Google에서 제공하는 Text-to-Speech API로, 텍스트 입력을 자연스러운 음성으로 변환한다. 본 시스템에서는 EEVE에서 생성한 응답 문장을 gTTS로 변환한 후 WAV 파일로 저장된다.

2.4 메타휴먼 캐릭터 모델링: Unreal Engine 블루프린트

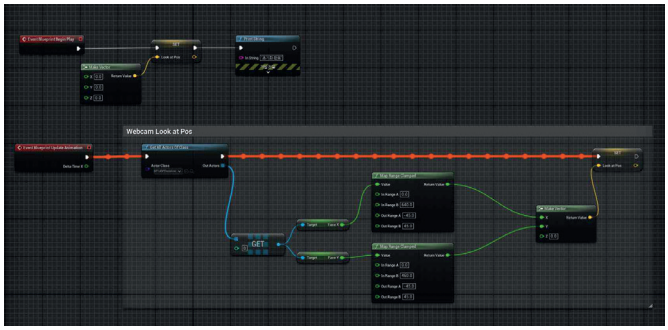

그림 2. Unreal Engine 이벤트그래프

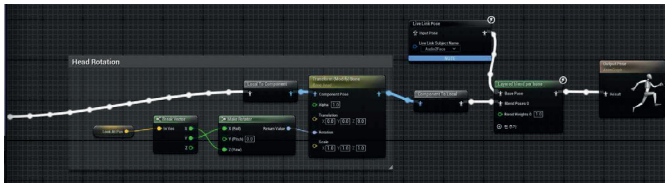

그림 3. Unreal Engine 애님그래프

애니메이션 적용 전, MetaHuman 캐릭터의 리깅과 제어 구조를 구성하기 위해 언리얼 엔진 블루프린트(Blueprint)에서 이벤트그래프와 애님그래프(AnimGraph)를 사용하였고 립싱크와 시선 움직임의 애니메이션이 충돌하지 않도록 슬롯 노드(Slot Node) 및 본 기반 레이어드 블렌딩(Layered Blend per Bone) 구조를 통해 얼굴 애니메이션을 정밀하게 제어하였다.

2.5 얼굴 애니메이션 생성: A2F, Unreal Engine의 LiveLink

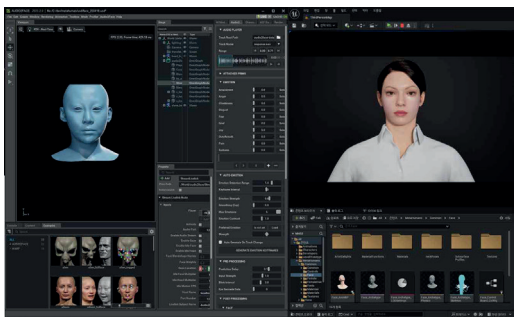

그림 4. A2F 및 Unreal Engine의 LiveLink 활성화

NVIDIA의 Omniverse 플랫폼에서 제공하는 AI 기반 얼굴 애니메이션 생성 도구로, 입력된 오디오의 음성 정보에 따라 Blend Shape 값을 자동 추출한다. REST API를 통해 Unreal Engine과 연동되며, 생성된 립싱크와 표정은 메타휴먼 캐릭터에 실시간으로 반영된다.

2.6. 구현 방안

그림 5. 현재 구현한 화면

Ⅲ. 결론

본 논문에서는 음성 인식부터 애니메이션 생성, 시선 추적까지 전 과정을 자동화한 대화형 메타휴먼 시스템을 설계하였다. 사용자의 음성과 얼굴 방향을 실시간으로 분석하고, 이에 맞춰 응답하고 반응하는 캐릭터를 통해 몰입감 있는 상호작용을 제공할 수 있음을 보였다. 본 시스템은 키오스크, 가상 상담, 교육 등 다양한 응용이 가능하며, 향후 감정 분석, 제스처 인식, 다중 사용자 대응 등 정밀한 기술을 추가함으로써 보다 정교하고 인간 중심적인 상호작용을 실현할 수 있을 것이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC)의 지원을 받아 수행된 연구임(IITP-2025-RS-2021-II211816)

참 고 문 헌

[1] Chojnowski, O., Eberhard, A., Schiffmann, M., Müller, A., and Richert, A., “Human-like nonverbal behavior with MetaHumans in real-world interaction studies: An architecture using generative methods and motion capture,” arXiv preprint, arXiv:2501.10713, 2025.

[2] Song, Y., and Xiong, W., “Large Language Model-Driven 3D Hyper-Realistic Interactive Intelligent Digital Human System,” Sensors, 25(6), pp. 1855, 2025.

[3] Kim, S., Choi, S., and Jeong, M., “Efficient and effective vocabulary expansion towards multilingual large language models,” arXiv preprint, arXiv:2402.14714, 2024.