

생성 AI와 음성 인식을 활용한 메타버스 가상 면접 플랫폼

윤채원*¹, 양성은*², 박재성*³, 시종욱**⁴, 정유철***⁵, 김성영****⁶

Metaverse Virtual Interview Platform Leveraging Generative AI and Speech Recognition

Chaewon Yoon*¹, Seongeun Yang*², Jaesung Park*³, Jongwook Si**⁴, Yuchul Jung***⁵, and
Sungyoung Kim****⁶

본 연구는 2024년도 과학기술정보통신부와 한국전파진흥협회의 메타버스랩지원사업에 의한 연구임

요약

본 논문에서 제안하는 새로운 가상 면접 플랫폼은 메타버스와 인공지능의 결합을 통해 구현된 시스템이다. 음성 생성 및 합성, GPT 기반의 질문 생성 기술을 활용하여, 사용자와 인터뷰 모델 간의 상호작용을 강화하고 감정이 담긴 음성 데이터를 통한 면접을 제공한다. GPT-4를 기반으로 하는 면접 질문 생성은 역량과 잠재 능력을 효과적으로 평가하기 위해 프롬프트 엔지니어링을 적용한다. 가상 면접 환경은 유니티 엔진을 통해 구축되며, 디지털 휴먼을 활용한 가상 면접관과 립싱크 기술의 적용으로 실제 면접과 유사한 상황을 보여준다. 제안 방법은 답변을 분석해 상세한 면접 결과를 제공하는 기능이 포함되기 때문에, 참여자들이 실제 면접 전에 자신의 능력을 점검하고 면접 기술을 향상시킬 수 있는 기회를 제공할 수 있다.

Abstract

This paper proposes a new virtual interviewing platform, implemented through the combination of the metaverse and artificial intelligence. Utilizing voice synthesis and generation, along with GPT-based question generation techniques, it enhances the interaction between users and the interview model and provides an interview experience enriched with emotion-laden voice data. The GPT-4 based interview question generation employs prompt engineering to effectively assess competencies and potential abilities. The virtual interviewing environment is built using the Unity engine and features digital human interviewers with lipsync technology, creating a scenario akin to real-life interviews. The proposed method includes a feature for analyzing responses to provide detailed interview results, thereby offering participants the opportunity to assess their capabilities and enhance their interview skills before the actual interview.

Keywords

virtual interview platform, metaverse, generative AI, speech, text

* 국립금오공과대학교 컴퓨터공학과 학사과정
- ORCID¹: <http://orcid.org/0009-0007-5871-5366>
- ORCID²: <http://orcid.org/0009-0008-2039-9916>
- ORCID³: <http://orcid.org/0009-0002-5431-2344>
** 국립금오공과대학교 컴퓨터·AI융합공학과 박사과정
- ORCID: <http://orcid.org/0000-0003-2092-2769>
*** 국립금오공과대학교 인공지능공학과 부교수
- ORCID: <http://orcid.org/0000-0002-8871-1979>

**** 국립금오공과대학교 컴퓨터공학과 교수(교신저자)
- ORCID: <http://orcid.org/0000-0002-7722-6759>

· Received: Mar. 07, 2024, Revised: Mar. 14, 2024, Accepted: Mar. 17, 2024
· Corresponding Author: Sungyoung Kim
Dept. of Computer Engineering, Kumoh National Institute of Technology,
61 Daehak-ro (yangho-dong), Gumi, Gyeongbuk, [39177] Korea
Tel.: +82-54-478-7530, Email: sykim@kumoh.ac.kr

I. 서 론

현대 사회는 지속적으로 변화하고 발전하면서 다양한 사회 문제와 경제적 도전에 직면하고 있다. 이러한 문제들에 효과적으로 대응하기 위해, 첨단 기술의 적용이 필수적이다. 특히, 코로나 팬데믹이 가져온 변화 속에서, 기업들은 구인난에 직면해 있으며, 구직자들은 취업의 어려움을 겪고 있다. 이 시점에서 메타버스와 AI 기술의 결합은 이러한 문제를 해결하는 데 있어 중요한 역할을 할 수 있다.

메타버스는 현실을 초월하는 가상의 우주를 의미하는 'Meta'와 'Universe'의 합성어로, 단순히 디지털 콘텐츠를 소비하는 차원을 넘어서 사회적, 문화적, 경제적 경험까지 가능케 하는 가상 세계의 진화된 형태를 말한다[1]. 특히 코로나 팬데믹 이후, 디지털 환경에 익숙한 세대가 확대됨에 따라, 메타버스 내에서의 소통과 상호작용이 증가하면서 큰 성장세를 보이고 있다. 이에 따라 메타버스와 관련한 연구들이 지속적으로 진행되고 있다[2-3].

이러한 메타버스의 부상은 4차 산업혁명의 진전과 코로나 팬데믹으로 인한 비대면, 디지털 중심의 사회 변화와 맞물려, 채용 시장에도 혁신을 가져왔다. 디지털 전환이 급속히 이루어지면서, 디지털 역량을 갖춘 인재에 대한 수요가 급증하고 있으며, 이는 2022년 채용 트렌드의 증가와 비대면 및 대면 채용 방식의 병행, 그리고 2023년 채용 시장 변화의 한 축으로 자리잡았다. 이런 변화에 발맞추어, 가상 인터뷰 시스템이라는 새로운 제안이 등장했다. 이 시스템은 메타버스의 상호작용적 특성을 활용하여 취업 준비생들이 실제 면접 환경과 유사한 조건에서 연습할 수 있게 함으로써, 면접에서 자기소개를 비롯한 여러 과제들을 능숙히 대처할 수 있도록 지원한다. 메타버스 환경에서 제공되는 가상 면접 시스템을 통해 취업 준비생들은 자신의 약점을 파악하고 보완할 수 있는 기회를 얻으며, 이를 통해 실제 면접에서 자신감을 키울 수 있다.

메타버스와 AI의 결합은 교육, 환경 보호 등과 같은 다양한 분야에도 긍정적인 영향을 미칠 수 있다. 예를 들어, 자연어 처리 및 컴퓨터 비전 기술을 이용하면, 교육 프로그램을 더 개인화하거나 의료 등의 분야에서 더 정밀하게 활용할 수 있다. 이를

통해 사회적 가치를 창출하고, 지속 가능한 발전을 도모할 수 있다.

메타버스 환경에서 구현되는 AI 기반 면접 플랫폼은 기업들에게 새로운 가능성을 제시한다. 이러한 플랫폼은 인재 채용 프로세스를 개선할 수 있는 잠재력을 지니고 있다. AI 기술을 활용함으로써, 기업들은 인재 선발 과정에서 발생할 수 있는 인적 오류를 줄이고, 더 정확하고 효율적인 방식으로 적합한 인재를 찾아낼 수 있다. 또한, 구직자들의 경우 이러한 플랫폼을 이용해 본인의 면접 능력을 향상시킬 수 있으며 신뢰있는 답변을 통해 면접을 대비할 수 있는 장점이 있다.

본 논문에서는 메타버스 환경에서의 다양한 기술의 융합을 통한 플랫폼을 소개하며 기술적인 측면과 기능에 대한 내용을 중점적으로 다룬다. 제안하는 기술은 생성 AI, GPT 기반 NLP, 음성 인식 및 합성 기술을 융합한 플랫폼에서 가상 인터뷰 서비스를 통해 효과적인 인터뷰를 실시할 수 있다.

2절에서는 AI 기반의 면접과 필요한 기술에 대한 관련 연구들을 소개하고 분석한다. 그리고 3절에서는 텍스트와 음성에 관련한 제안 방법, 4절에서는 3절의 내용을 메타버스 환경에 적용하는 기술에 대해 설명한다. 마지막으로 5절에서는 결론과 향후 과제에 대하여 소개한다.

II. 관련 연구

2.1 AI 기반 면접

생성 AI는 기존 데이터를 기반으로 비교 학습을 통해 새로운 콘텐츠를 생성하는 인공지능 기술을 말한다. 이러한 생성형 AI 시스템은 맞춤형 교육 및 콘텐츠 제작, 데이터 증강 등 다양한 분야에서 무한한 가능성을 제공한다[4-5]. 또한, AI 면접 분야에서도 활발히 사용된다.

안정적인 AI 면접 서비스를 제공하기 위해, aSSIST 연구팀은 효율성 및 공정성을 확보하는 면접 시스템 도입을 위해 40만개 이상의 인터뷰 영상과 10만개 이상의 평가 데이터 세트를 구축하여 모델을 학습하였다[6].

이 연구는 서류평가 단계에서 신규 지원자의 이력서를 경력직 지원자의 이력서와 비교하고 유사성을 판단할 수 있는 모델을 구축했다. 또한, 직무 요건을 정의한 후 해당 요건에 기반한 질문을 자동 생성할 수 있는 생성형 AI를 사용하였다. 이후 AI는 지원자의 직무경력과 역량 등의 정보를 종합 분석하여 지원한 직무와 가장 적합할 것으로 보이는 후보자를 추천하는 평가 방식을 사용했다. 제시된 방안은 영상 데이터를 구축하고 추천 시스템으로 후보자를 평가하는 방식에서 차별화를 둔다는 점에서 제안 방법과의 차이가 있지만, 생성형 AI를 활용해 질문을 생성한다는 공통점이 있다.

J. Achiam et al.은 GPT-4[7]를 사용해 질문 생성 조건 지정으로 예/아니오의 형식이 아닌 ‘어떻게, 무엇을, 왜’에 대한 답변을 얻을 수 있는 개방형 방식의 질문을 생성함으로써 질문의 깊이를 조절하였다[8]. 또한, 반복적인 질문 생성을 피하기 위해 이전 대화 기록을 저장하여 현재 대화 기록과 다른 새로운 주제의 질문을 생성할 수 있도록 하였다. 그리고 이전 질문과 생성된 질문을 비교하여 맥락에 맞지 않는 답변을 할 경우, 해당 질문을 제거한 후 답변 재생성 기회를 주어 생성 모델이 기존에 부여한 역할에서 벗어난 질문을 생성하지 않도록 중재하였다. 이 연구는 GPT 모델을 활용해 질문을 생성하는 점에서 제안된 방식과 유사하지만, 다양한 상황에 대응하는 추가 기능이 포함된 것이 특징이다. 반면, 제안된 방법은 실제 면접 상황에서 나올 법한 질문에 중점을 둔다. 따라서, 맥락에 맞지 않는 응답에는 낮은 점수를 부여하며, 꼬리 질문을 통해 실제 면접과 비슷한 경험을 제공한다.

2.2 립싱크 기술

유니티에서 활용 가능한 립싱크 기술은 여러 종류가 존재한다.

SALSA Lip-sync[9]는 2D, 3D 캐릭터에 간단한 자동화된 립싱크 기능과 고품질의 립 동기화를 구현해낼 수 있다. 이 기술은 음소 매핑이나 베이킹 없이 오디오 입력에 따라 립싱크를 생성해낼 수 있는 실시간 시스템이다. 또한, 다양한 기술을 사용하여 오디오 파형 분석을 기반으로 하여 립싱크의 자동

화를 제공한다. 그리고 디지털 휴먼 모델에 적용하여 립싱크 기능을 작동하게 할 수 있다[10]. 하지만, 가격 측면에서는 효율적이지 못하다는 단점이 있다.

Lipsync pro[11]는 음소 기반의 립싱크와 얼굴 애니메이션 시스템을 제공하는 도구이다. 이 기술은 BlendShape이나 뼈대로 구성된 음소 및 감정에 대한 얼굴 감정 표현을 손쉽게 작업하고 음소 동기화를 할 수 있다[12].

Character Generator[13]는 3D 모델 디자인 및 애니메이션 기술을 갖춘 게임 엔진용 3D 캐릭터를 생성할 수 있는 기능을 지닌다. 두 개의 캐릭터를 블렌딩 하여 3D 캐릭터를 생성할 수 있으며 캐릭터 커스터마이징, 오디오 소스에 대한 립싱크 기능을 가지고 있다. 하지만, Lipsync Pro[11]와 Character Generator[13]는 현재 지원이 중단된 상태이기 때문에 활용하기에 어려움이 있다.

uLipsync[14]는 실시간으로 오디오 입력을 분석하여 3D 캐릭터 모델의 입 모양을 동기화하는 기술이다. WAV 파일 형식의 오디오 데이터와 입모양이 실시간으로 동기화되며, 오디오 데이터를 분석하여 특정 음성 주파수를 감지할 수 있다. 따라서, 파일을 통해 실시간으로 오디오를 처리하여 립싱크를 할 수 있다는 것이 가장 큰 특징이다. 따라서, 본 논문에서는 uLipsync를 이용하여 가상 면접관에 립싱크 기술을 적용한다.

2.3 음성 처리

사용자 경험은 다양한 감각과 관련된 요소들의 영향을 크게 받으므로 시각적, 청각적 관련 모델이 사용자의 경험을 결정하는 주요 원인이 된다[15].

STT(Speech-to-Text)는 사용자의 말을 인식하여 모델에게 텍스트 형식으로 반환하는 기술이다. 모델의 질문 생성이나 답변 생성에 중요한 영향을 끼치기 때문에 정확도가 매우 중요하다. 그러나 소요시간 단축 또한 사용자 경험을 높이는 중요한 요소이기 때문에 너무 무거운 모델은 오히려 사용자의 몰입감만 낮추는 요소로 작동한다. 모델의 정확도와 소요시간 사이의 Trade-off를 균형있게 조절하여 효율성을 극대화하는 것이 중요하다.

TTS(Text-to-Speech)는 사용자에게 음성을 제공하는 만큼 몰입도에 중요한 영향을 미친다. 또한, 별도로 보코더를 이용하는 방식과 End-to-End 방식 간의 추론속도의 차이가 존재하기 때문에 추론속도와 안정성을 모두 고려해서 적합한 TTS를 선정하는 것이 중요하다.

대표적인 STT와 TTS에 대한 연구는 다음과 같다. Meta사의 wav2vec2[16]는 Pre-training 과정에서 마스킹 기법을 사용한다. 그리고 End-to-End방식으로 Transformer 블록을 활용하며, Diversity loss이용하여 Multiple Codebook을 효과적으로 사용한 STT이다.

Tacotron2[17]는 Attention을 기반으로 하여 Seq-to-Seq TTS 모델 구조를 사용한 End-to-End TTS를 제안하였다. 그리고 플로우 기반 생성 모델과 동적 프로그래밍의 속성을 활용하여 MAS(Monotonic Alignment Search)기법을 이용한 카카오 사의 Glow-TTS[18], Glow-TTS의 MAS를 사용하고 Diffusion model의 개념을 사용한 score 기반 디코더로 구성된 Grad-TTS[19], Variational Inference와 GAN을 결합한 방식을 사용하는 End-to-End 방식의 TTS인 VITS[20] 등이 존재한다. TEA-VITS[21]은 시간별 감정에 따른 자동 감정 라벨링과 End-to-End방식의 TTS인 VITS를 결합한 TTS이다. 이러한 방식

은 자연스러운 발음, 감정 표현, 억양 등 사람의 음성 에 가까운 결과물을 생성하는데 효과적이다.

III. 음성 기반 면접 질문 생성 및 응답

본 논문에서 소개하는 가상 면접 플랫폼의 전체적인 과정은 그림 1과 같다.

메타버스 환경에서 사용자가 가상 면접을 시작하면, 서버와 사용자는 Python 기반의 TCP 소켓 통신을 사용하여 통신한다. 면접 질문은 사용자로부터 시작 트리거 문자를 받으면 OpenAI의 GPT-4 모델을 기반으로 면접 질문이 생성된다. 생성된 질문 텍스트는 사용자가 질문 텍스트를 음성으로 들을 수 있도록 TTS를 통해 음성으로 변환되어 사용자에게 전달되며 사용자는 해당 질문을 듣고 다시 서버에 답변을 보낸다. 이때, STT를 통해 사용자의 음성은 텍스트로 변환이 되며, 사용자의 답변 텍스트 모델에 입력되어 새로운 면접 질문을 생성하는 구조로 이루어진다. 이후 사용자로부터 면접 종료 및 답변 평가를 위한 평가 트리거 문자를 받으면 OpenAI의 GPT-4 모델을 통해 지금까지 진행한 면접 질의응답 정보를 토대로 면접 답변에 대한 평가 분석 보고서가 생성된다.

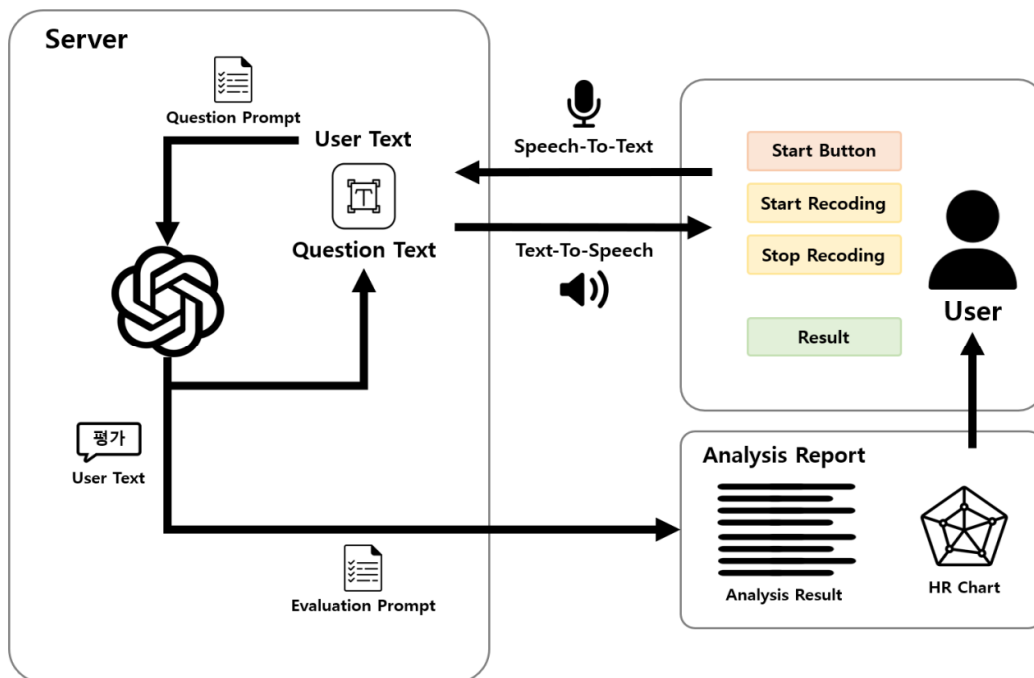


그림 1. 제안하는 방법의 전반적인 흐름도
 Fig. 1. Overall flowchart of proposed method

3.1 음성 처리

음성처리 기술은 사람의 음성을 텍스트로 변환하는 기술인 STT와 텍스트를 음성으로 추론하여 원하는 목소리로 출력하는 기술인 TTS가 사용되었고 이는 사용자와 인터뷰 모델의 상호작용이 원활하도록 돕고 사용자가 면접관과 실제로 대화하는 듯한 현실감을 증폭하는데 큰 역할을 한다.

본 논문에 사용된 STT모델은 OpenAI사의 Whisper API, TTS모델은 VITS이다. STT모델은 각 언어마다 WER과 CER이 다른데, Whisper API의 경우 다국어 지원뿐만 아니라 한국어의 오류율이 비교적 낮은 것을 알 수 있다. 면접관은 말에는 감정이 최대한 배제되어 있다. 그러나 압박 면접과 같은 특수한 면접 상황에서는 말에 감정이 담겨야 하는 상황이 존재한다. 우리는 이런 문제를 해결하기 위하여 Speech Emotion Diarization[22]을 사용한다. 이를 통해 각 순간별 음성의 감정을 확인하고 이를 토대로 계산하여 음성의 전체적인 감정을 라벨링하는 기법을 고안한다. 이는 기존 감정 인식 모델의 VAD[23] 측정에 단어가 영향을 미치는 정도를 최소화하여 감정은 기쁘지만 음성의 피치는 낮고, 반대로 슬프지만 피치는 높은 예외적인 상황을 최소화하는 것을 방안으로 안정적이고 직관적인 감정 라벨링을 실행한다.

3.2 면접 질문 생성

사용자 고유의 잠재 능력과 역량, 인성, 지식을 효과적으로 평가하기 위해서는 면접 질문의 품질이 중요하다. 면접자의 능력과 장점을 충분히 드러낼 수 있도록 유도하는 질문을 생성하는 것은 이후 면접 답변 분석 시 AI가 답변을 통해 지원자의 지식의 정도와 사고방식, 행동방식, 가치관 등을 이해하는데 도움을 준다. 따라서 생성형 AI가 지원자를 평가할 면접 질문을 생성할 경우 지원자의 역량과 잠재 능력을 평가할 수 있도록 설계되어야 한다.

본 논문에서는 심층적인 면접 질문을 생성하기 위해 OpenAI의 GPT-4기반 모델의 Prompt Engineering을 통해 면접 질문 생성 조건을 조정한다. 생성 AI 모델의 역할을 엄격하고 진지한 대화

를 진행하는 면접관으로 설정하여 모델이 면접 질문 이외의 답변을 생성하지 못하도록 제한한다. OpenAI GPT-4 모델의 입력값은 사용자의 답변 음성이 STT 과정을 거친 텍스트 값이므로 발음이나 문맥이 매끄럽지 않은 경우가 빈번히 존재한다. 따라서, AI 모델이 사용자의 답변을 보다 정확하게 이해할 수 있도록 모델의 입력값이 음성인식으로 생성된 값임을 명시해준다. 실제 면접 과정과 유사하게 진행하기 위해 한가지 질문에 대한 사용자의 답변을 획득한다. 이 답변을 기반으로 생성형 AI가 더 심층적인 질문을 할 것인지, 다른 주제의 질문을 할 것인지를 판단하여 지원자에게 새로운 면접 질문을 제시하도록 질문 생성 시 하나의 질문만을 생성하도록 설계한다. 만약, 이때 사용자가 면접 질문에 제대로 된 답변을 하지 못하거나 질문과 무관한 답변을 할 경우에 모델은 사용자에게 답변의 재기회를 주거나 현재 질문과 다른 새로운 질문을 제공한다. 또한, 트리거 문자인 “시작”으로 설정하여 해당 트리거 문자를 전달받을 경우 저장된 질의응답 정보가 초기화되고 새로운 면접이 진행될 수 있도록 한다. 이러한 Prompt Engineering 과정을 통해 사용자는 더 심층적이고 맞춤형 면접 질문을 생성받을 수 있다. 그림 2는 위의 Prompt Engineering을 통해 질문을 생성하여 사용자와의 면접 진행 과정의 예를 보여준다.

3.3 면접 분석 보고서 생성

AI 면접에서 생성 AI가 지원자의 답변에 대한 분석 결과를 통해 채용자는 지원자의 잠재 능력과 역량, 인성, 지식의 정도를 파악할 수 있다. 따라서 생성 AI는 지원자의 답변에 대한 단순한 분석 너머에 있는 지원자의 사고방식, 가치관, 역량 등에 대한 분석을 할 수 있도록 지원자의 답변에 대한 의도를 잘 이해하고 있어야 한다. 또한, 면접 평가 시 공정성과 신뢰성은 매우 중요한 요소이다. 평가 요소가 특정 기준에 편향되어 있다면 면접자의 역량을 구체적으로 평가하기 어려워지며 기준에 대한 명확한 정의없이 평가하게 될 경우, 일관성 없는 평가 결과로 인해 편향되거나 공정하지 않은 분석이 이루어질 가능성이 높다.

```

>>Question : 안녕하세요, 저는 이번 인터뷰를 진행할 담당자입니다. 먼저 간단하게 자신을 소개해 주시겠어요?
>>Answer : 안녕하세요 저는 금호공과대학교 3학년에 재학 중인 박재성입니다. 저의 주 전공으로는 음성 처리이고, 부 전공으로는 자연화 처리를 맡고 있습니다. 잘 부탁드립니다.
>>Question : 박재성 님, 안녕하세요. 여기에 참석해주셔서 감사합니다. 음성 처리와 자연어 처리에 대한 전공 경험이 있더니 매우 흥미롭군요. 그럼, 이 전공을 선택한 이유와 이를 통해 어떤 역량을 갖추었는지 자세히 설명해주실 수 있으신가요?
>>Answer : 제 전공과 관련된 가장 인상깊었던 경험으로는 가상 현전 모델 제작이었던 것 같은데요. 거기에서 저는 좀 더 정확한 음성 2배 조금 더 자연스러운 음성 바랄 모델 제작에 주로 들어왔고 GPT 같은 경우 GPT 모델을 음성으로 바꾸는 그런 쪽에 조금 더 신경을 썼던 것 같습니다.
>>Question : 그 경험을 통해 어떤 기술적인 문제를 해결했는지, 그리고 그 과정에서 어떤 역량을 발휘했는지 구체적으로 설명해주시겠어요?
>>Answer : 그 프로젝트를 진행하면서 저는 모델 구조에 대해서 조금 어려움에 봉착했었는데요. 그 뿐만 아니라 음성적이다 보니까 데이터가 조금 더 부족했던 점, 그런 점을 극복하기 위해서 저는 전문가에게 자문을 구해서 해결했던 경험이 있고요. 그 과정에서 저는 문제 해결 능력, 그러니까 상대처 능력을 조금 더 발의했다고 생각합니다.
    
```

그림 2. 프롬프트 엔지니어링을 통한 면접 질문 생성 및 면접 과정의 예

Fig. 2. Generating interview questions through prompt engineering and examples of the interviewing process

이는 평가에 대한 신뢰도를 바탕으로 답변에 대한 보완점을 얻고자 하는 사용자에게 실질적인 정보를 제공하지 못하고 서비스의 가치를 하락시키는 결과를 불러올 수 있다.

사용자에게 면접 평가 과정에 대한 공정성과 신뢰성을 보장하기 위해 면접 평가 요소는 특정 기준에 편향되지 않도록 주의해야하며, 명확한 평가 기준을 정의하여 공정하고 신뢰도 높은 평가 분석 자료를 제공하는 것이 중요하다. 본 논문에서는 면접 분석 보고서가 일관성 있는 평가와 정밀한 분석이 가능하도록, GPT-4기반 모델의 Prompt Engineering을 통해 면접 분석 조건을 조정하였다. 생성 AI는 사용자로부터 “평가”라는 트리거 문자를 받으면 누적된 면접 질의응답 정보를 바탕으로 지원자를 평가하며, 이에 대해 평가 점수를 부여한 이유를 설명하고 지원자의 답변에서 보완점 또는 개선점 등 특정 부분에 대한 조언을 면밀히 분석한다. 현재 사용자의 답변에 대한 평가 요소는 총 5가지로 지원자가 면접 질문에 적절한 답변했는지, 자신의 경험에 빗대어 적절한 주장을 하고 있는지, 불필요한 답변을 하고 있는지를 포함한 답변 응답 분석을 진행한다. 답변의 분량이 짧거나 시간을 초과하지 않는지를 평가하는 분량의 적절성과 발음을 명확하게 하여 답변을 했는가를 평가하는 발음의 명확성에 대한 평가 분석은 필수적으로 포함된다[24]. 평가 채점 척도는 “매우 좋음”부터 “매우 나쁨”까지 5가지 척도로 나뉘며 기본 점수를 “매우 나쁨”으로 설정하여 지원자의 답변에 따라 점수가 가산되는 형태로 평가가 이루어진다.

면접 분석이 모두 끝나면 면접 답변에 대한 10가지 역량 점수가 매겨진다. 최소 0점에서 최대 10점으로 점수가 부여되며 이는 사용자에게 보여지는 평가 보고서 중 역량 그래프에 포함될 값으로 지원

자의 면접 답변을 통해 확인할 수 있는 상위 5가지의 역량 값을 그래프로 보여준다. OpenAI의 GPT-4 모델을 통해 얻은 10가지의 HR 역량 점수를 추출하여 JSON 형태로 값을 변경한 후 점수가 높은 상위 5개의 값을 저장하여 면접 답변 분석 결과와 함께 JSON 형태로 사용자에게 전달한다.

3.4 면접 질문 및 답변 Instruction set 구축

특정 도메인에 특화된 모델을 만들기 위해서는 해당 도메인과 연관된 데이터로 이루어진 데이터셋의 구축이 필요하다. 이렇게 학습된 모델을 사용하여 질의를 할 경우 원하는 방식의 답변을 손쉽게 얻을 수 있으며 학습 전보다 질문의 흐름을 더 잘 이해하여 보다 신뢰도 높은 답변을 도출할 수 있다. Instruction set은 모델이 학습 데이터를 어떻게 해석하고, 어떤 유형에 대해 어떤 방식으로 대응해야 하는지에 대한 길잡이 역할을 하기 때문에 모델 학습 시 Instruction set의 구성은 성능 향상에 매우 중요한 역할을 한다. AI 인터뷰 모델은 지원자의 능력을 평가하기 위해, 다양한 질문과 답변 데이터에 대한 이해와 특정 직군에 대한 충분한 지식을 가지고 있어야 한다.

또한, 면접 과정에서 지원자의 잠재 능력 및 역량, 직무 적합성을 파악하기 위해서는 단순히 기술적 이해를 위한 질문 뿐만 아니라 지원자의 인적성을 평가할 수 있는 질문도 필요하다. 생성형 AI를 사용하여 지원자를 면밀히 평가할 수 있는 질문을 효과적으로 생성하기 위해 반자동 Instruction set을 구축하였다. HR 항목 분석 과정에서 HR 항목 유형으로 M. M. Robles의 “Executive Perception of the Top 10 Soft Skills Needed in Today’s Workplace”[25]의 10가지 HR 역량을 활용한다.

IV. 메타버스 환경에서의 가상 면접 적용

3절에서 언급한 면접 질문 생성 및 분석 기술을 메타버스 환경에서 서비스를 제공하며, 가상 면접은 크게 면접장과 면접 과정으로 나누어진다. 가상 면접장은 유니티 엔진을 사용하여 구성하며, 이를 통해 가상 면접을 진행할 수 있도록 한다[26]. 또한, 다양한 회사나 학교에 알맞게 면접 상황 및 환경을 구축하여 다양한 면접을 경험할 수 있도록 한다.

가상 면접 플랫폼을 시작하면, 그림 3과 같이 원하는 면접을 선택할 수 있는 초기 화면이 나타난다. 이를 통해 사용자는 원하는 기업을 선택하여 해당 면접장으로 이동하여 면접을 진행할 수 있다.

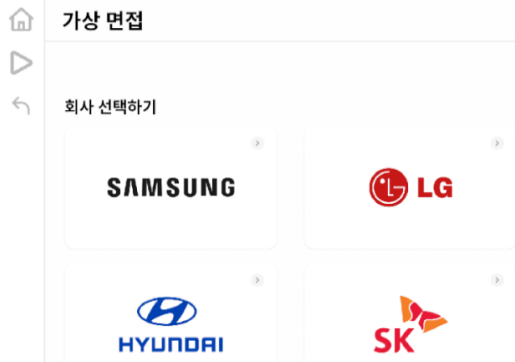


그림 3. 가상 면접 플랫폼의 초기 화면
Fig. 3. Initial scene of virtual interviewing platform

사용한 가상 면접장은 Business Office[27]이며, 그림 4와 같다. 면접장은 책상, 의자, 노트북 등의 에셋을 활용해 구성한다. 그리고 조명을 적절히 배치하여 면접장을 더욱 현실감 있고 자연스럽게 구성한다. 면접장은 하나의 층으로 구성되며, Lighting Settings Assets로 Office Lighting Setting을 적용하여 도심에 있는 듯한 느낌을 준다.



그림 4. 에셋을 활용한 가상 면접장 구축
Fig. 4. Building virtual interview room using asset

면접장에서 면접을 진행하는 주체인 가상 면접관은 디지털 휴먼을 제작하여 사용한다. 디지털 휴먼이란 고도의 그래픽 기술을 활용해 인간의 외형을 현실적으로 시각화한 기술이다. 이는 인간과 유사한 외모와 행동을 통해 사용자들에게 자연스럽고 현실적인 경험을 제공할 수 있다. 따라서, 본 논문에서는 가상 면접관으로 사용하기 위해 디지털 휴먼을 제작한다.

본 논문에서는 생성된 음성에 대한 립싱크를 위해 uLipsync[14]를 가상 면접관에 적용하며 과정은 다음과 같다. 분석된 오디오 데이터에서 'a', 'e', 'i', 'o', 'u'와 같은 모음이나 'b', 'm', 'p'와 같은 자음과 같은 음소를 식별한다. 그리고 식별된 음소를 특정 입 모양에 매핑하며, 캐릭터의 입술 애니메이션을 생성한다. 이를 오디오 플레이백과 정확하게 동기화하여, 면접관이 말하는 것처럼 보이도록 한다.

그림 5는 앞에서 언급한 모음인 'a', 'e', 'i', 'o', 'u'를 면접관에 적용한 얼굴을 나타낸다. 가상 면접관에 립싱크 기능을 적용함으로써, 자연스러운 대화를 가능하게 하고 사용자가 면접에서 몰입감을 느낄 수 있다.



그림 5. 가상 면접관의 모음 적용 결과 (a, e, i, o, u)
Fig. 5. Results of vowel application in virtual interviewer (a, e, i, o, u)

제안하는 플랫폼에서 사용자는 키보드를 통해 면접장에서 이동할 수 있으며 의자에 앉으면 면접이 시작된다. 그림 6은 면접 시작 시 초기 화면을 나타내며 3가지 버튼으로 구성된다. Start Button은 “시작”이라고 녹음된 WAV 파일을 서버로 전송하는 기능을 한다. 이를 통해 서버에서는 시작입을 인지하고 면접관이 질문을 시작하게 된다. Start Recording은 마이크를 통해 입력으로 받은 음성을 녹음하는 기능을 담당한다. 그리고 Stop Recording은 녹음을 중단하고 WAV 파일을 생성하며, 이 파일을 서버로 전송하는 역할을 한다.



그림 6. 면접 시작 시 초기 화면
Fig. 6. Initial scene at the start of interviewing

이렇게 녹음한 WAV 파일을 서버에 전송하면 서버에서 STT와 질문 생성, TTS 과정을 거쳐 WAV 파일로 전달받는다. 이어서 이 파일을 오디오 클립

으로 변환하며, 면접관의 Audio Source 오디오 클립으로 할당한다. 따라서, 면접관이 해당 WAV 파일에 맞게 입모양이 변화할 수 있다.

질문과 답변을 3번 이상 진행하면 결과를 추론할 수 있는 버튼이 활성화되어, 사용자가 원하는 시기에 면접을 종료할 수 있다. 면접을 종료하게 되면, 답변에 대한 평가가 이루어질 수 있는 씬으로 이동하게 된다.

그림 7은 면접 결과와 분석 내용이 Scene에 출력된 예시를 나타낸다.

면접 답변 평가 항목으로 5가지 항목에 대하여 평가한다. 질문에 적절한 대답을 하였는지, 자신의 경험에 빗대어 적절한 주장을 하고 있는지, 불필요한 답변을 하고 있는지, 답변의 분량이 적절한지 혹은 시간을 초과하지 않았는지, 명확한 발음이나 목소리톤 그리고 말하기 속도가 적절한지에 대해 10점 만점에 대해 평가값을 보여주며, 각각의 항목에 대해 피드백을 제공한다. 총평으로 전체 면접 결과를 요약하여 답변에서의 긍정적인 점과 보완해야 할 부분을 구체적으로 명시해준다.

문장으로 면접을 분석해줄 뿐 아니라, 레이더 차트[28]를 이용하여 의사소통, 대인관계능력, 진실성, 적응성, 팀워크 항목에 대한 분석 결과도 제공한다. 분석 결과는 서버에서 JSON 파일을 제공받아, 레이더 차트의 값을 표시하는 방식이다.

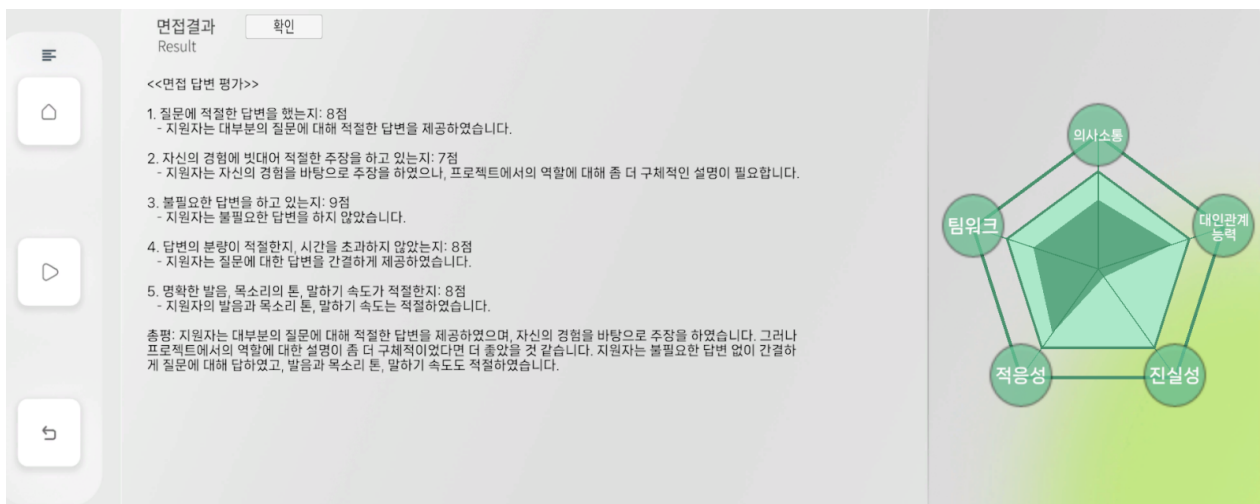


그림 7. 면접 결과 분석 예시
Fig. 7. Example of interview result analysis

V. 결론 및 향후 과제

본 연구에서는 메타버스와 인공지능(AI)을 결합한 새로운 가상 면접 플랫폼을 제안하였다. 이 플랫폼은 메타버스 환경에서 음성 생성 및 합성, GPT 기반의 질문 생성 기술을 활용하여 사용자 경험과 효율성을 높이는 것을 목표로 한다. 이를 위해 GPT-4 모델을 파인튜닝하여 면접 질문을 생성하도록 하고, 메타버스 플랫폼에서 음성 기술과 결합하여 실제와 같은 면접을 가능하게 하였다.

그러나, 이 플랫폼은 면접 질문에 대한 답변과 분석 결과의 신빙성에 대한 추가적인 고려가 필요하다. 특히, 실제 면접관의 분석 및 면접자의 비언어적 신호, 예를 들어 몸짓이나 동공 움직임 등을 포함하는 분석이 필요하다. 향후 연구에서는 실제 면접관의 평가와 비언어적 소통을 위한 햅틱 기술과 같은 추가적인 기술들을 통합할 계획이다.

References

- [1] N. G. Narin, "A Content Analysis of the Metaverse Articles", *Journal of Metaverse*, Vol. 1, No. 1, pp. 17-24, Dec. 2021. <https://orcid.org/0000-0002-4840-5408>.
- [2] K. Ryu, K. Park, S. Cho, and D. Kim, "A Study on the Application Direction of Financial Industry Metaverse Platform to secure MZ Generation Contact Points", *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol. 23, No. 3, pp. 127-137, Jun. 2023. <https://doi.org/10.7236/JIIBC.2023.23.3.127>.
- [3] Y. Zhen, K. Gwak, and Y. Rho, "A Study on Metaverse Framework Design for Education and Training of Hydrogen Fuel Cell Engineers", *The Journal of The Institute of Internet, Broadcasting and Communication*, Vol. 24, No. 1, pp. 207-212, Feb. 2024. <https://doi.org/10.7236/JIIBC.2024.24.1.207>.
- [4] J. Si, S. Yang, D. Kim, and S. Kim, "Metaverse Interview Room Creation with Virtual Interviewer Generation Using Diffusion Model", *IEEE Asia-Pacific Conference on Computer Science and Data Engineering*, Nadi, Fiji, pp. 1-4, Dec. 2023. <https://doi.ieeecomputersociety.org/10.1109/CSDE59766.2023.10487677>.
- [5] J. Si, J. Song, M. Woo, D. Kim, Y. Lee, and S. Kim, "Generative AI models for virtual interviewers: applicability and performance comparison", *International Conference on Innovation, Communication and Engineering*, pp. 27-28, Nov. 2023. <https://doi.org/10.1049/icp.2024.0193>.
- [6] B. C. Lee and B. Y. Kim, "Development of an AI-based interview system for remote hiring", *International Journal of Advanced Research in Engineering and Technology*, Vol. 12, No. 3, pp. 654-663, Mar. 2021. <http://dx.doi.org/10.34218/IJARET.12.3.2021.060>.
- [7] J. Achiam, et al., "GPT-4 Technical Report", arXiv:2303.08774, Mar. 2023. <https://doi.org/10.48550/arXiv.2303.08774>.
- [8] F. Chopra and I. Haaland, "Conducting Qualitative Interviews with AI", *CESifo Working Paper*, No. 10666, pp. 1-74, Sep. 2023. <http://dx.doi.org/10.2139/ssrn.4583756>.
- [9] SALSA Lipsync suite, <https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442> [accessed: Nov. 11, 2023]
- [10] J. Kum and M. Lee, "Can Gestural Filler Reduce User-Perceived Latency in Conversation with Digital Humans", *Applied Sciences*, Vol. 12, No. 21, pp. 10972-10988, Oct. 2022. <https://doi.org/10.3390/app122110972>.
- [11] Lipsync pro, <https://lipsync.rogodigital.com/> [accessed: Nov. 11, 2023]
- [12] X. Huang, J. Twycross, and F. Wild, "A Process for the Semi-Automated Generation of Life-Sized, Interactive 3D Character Models for Holographic Projection", *International Conference on 3D Immersion (IC3D)*, Brussels, Belgium, pp. 1-8,

- Dec. 2019. <https://doi.org/10.1109/IC3D48390.2019.8975993>.
- [13] Character creator, <https://www.reallusion.com/character-creator/> [accessed: Nov. 11, 2023]
- [14] uLipSync, <https://github.com/hecomi/uLipSync> [accessed: Nov. 11, 2023]
- [15] J. Lee and J. Kim, "A Study on Spatial Characteristics of Immersion and Reality in Cases of VR and AR Technology and Contents", *Journal of the Korean Institute of Interior Design*, Vol. 28, No. 3, pp. 23, Jun. 2019. <https://doi.org/10.14774/JKIID.2019.28.3.013>.
- [16] A. Baeviski, et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", *Advances in neural information processing systems*, pp. 12449-12460, Oct. 2020.
- [17] J. Shen, et al., "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Calgary, AB, Canada, pp. 4799-4783, Apr. 2018. <https://doi.org/10.1109/ICASSP.2018.8461368>.
- [18] J. Kim, et al., "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search", *Advances in Neural Information Processing Systems*, Vol. 33, pp. 8067-8077, Oct. 2020.
- [19] V. Popov, et al., "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech", *International Conference on Machine Learning*, pp. 8599-8608, Jul. 2021.
- [20] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech", *International Conference on Machine Learning*, Vol. 139, pp. 5530-5540, Jul. 2021.
- [21] J. Park, S. Son, and Y. Jung, "TEA-VITS : Emotion Voice Synthesis", *International Conference on Innovation, Communication and Engineering*, Nov. 2023. <https://doi.org/10.1049/icp.2024.0235>.
- [22] Y. Wang, M. Ravanelli, and A. Yacoubi, "Speech Emotion Diarization: Which Emotion Appears When?", *IEEE Automatic Speech Recognition and Understanding Workshop*, Taipei, Taiwan, pp. 1-7, Dec. 2023. <https://doi.org/10.1109/ASRU57964.2023.10389790>.
- [23] S. Park, et al., "Dimensional Emotion Detection from Categorical Emotion", *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4367-4380, Nov. 2021. <https://doi.org/10.18653/v1/2021.emnlp-main.358>.
- [24] C. Yoon, S. Park, and Y. Jung, "Constructing a Semi-automatic Instruction Set for Effective AI interview Models", *Proceedings of KIIT Conference*, Jeju, Korea, pp. 408-410, Jun. 2023.
- [25] M. M. Robles, "Executive Perception of the Top 10 Soft Skills Needed in Today's Workplace", *Business communication quarterly*, Vol. 75, No. 4, pp. 453-465, Oct. 2012. <https://doi.org/10.1177/1080569912460400>.
- [26] S. Yang, J. Si, and S. Kim, "Development of a Unity-Based Interview room Prototype for Virtual Interviews in a Metaverse Environment", In *Proceedings of Korea Society of Computer and Information Conference*, Vol. 31, No. 2, pp. 225-226, Jul. 2023.
- [27] Business Office, <https://assetstore.unity.com/packages/3d/environments/business-office-235285> [accessed: Nov. 11, 2023]
- [28] Radar Chart, <https://assetstore.unity.com/packages/tools/gui/radar-chart-50074> [accessed: Nov. 11, 2023]

저자소개

윤 채 원 (Chaewon Yoon)



2021년 3월 ~ 현재 :
국립금오공과대학교
컴퓨터공학과 학사과정
관심분야 : 자연어처리, 감성/성향
분석, GPT

양 성 은 (Seungeun Yang)



2020년 3월 ~ 현재 :
국립금오공과대학교
컴퓨터공학과 학사과정
관심분야 : 메타버스, 가상현실,
증강현실

박 재 성 (Jaesung Park)



2019년 3월 ~ 현재 :
국립금오공과 대학교
컴퓨터공학과 학사과정
관심분야 : 음성인식, 음성합성,
OCR

시 종 욱 (Jongwook Si)



2020년 8월 : 국립금오공과대학교
컴퓨터공학과(공학사)
2022년 2월 : 국립금오공과대학교
컴퓨터공학과(공학석사)
2022년 3월 ~ 현재 :
국립금오공과대학교 컴퓨터·AI
융합공학과 대학원(박사과정수료)

2023년 9월 ~ 현재 : 국립금오공과대학교 인공지능공학과
강사

관심분야 : 영상처리, 컴퓨터비전, 인공지능, 이상 감지,
영상 기반 생성 AI

정 유 철 (Yuchul Jung)



2011년 2월 : 한국과학기술원
(KAIST) 전산학과(공학박사)
2009년 1월 ~ 2013년 7월 :
한국전자통신연구원 (ETRI)
선임연구원
2013년 8월 ~ 2017년 8월 :
한국과학기술정보연구원 (KISTI)

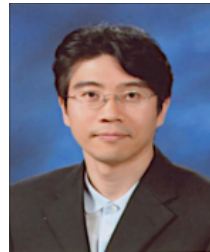
선임연구원

2017년 9월 ~ 2022년 2월 : 국립금오공과대학교
컴퓨터공학과 조교수

2022년 2월 ~ 현재 : 국립금오공과대학교 인공지능공학과
부교수

관심분야 : 정보 검색, 자연어 처리, 음성인식, 음성합성,
인공지능 기술 기반 응용

김 성 영 (Sungyoung Kim)



1994년 2월 : 부산대학교
컴퓨터공학과(공학사)
1996년 2월 : 부산대학교
컴퓨터공학과(공학석사)
2003년 8월 : 부산대학교
컴퓨터공학과(공학박사)
2004년 ~ 현재 :

국립금오공과대학교 컴퓨터공학과 교수

관심분야 : 영상처리, 컴퓨터비전, 기계학습, 딥러닝,
메타버스