JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 24, No. 5, pp. 999-1009, May. 2023

Check for updates

영어 초보자의 영어 학습 능력 향상을 위한 딥러닝 기반 가상 튜터 프로그램

주 재 성[1] · 홍 우 진[1] · 문 우 혁[1] · 유 사 라[1] · 고 경 수[1] · 고 영 서[1] · 장 하 은[1] · 김 시 현[1] · 신 정 훈[2] · 최 승 호[3*]
[1]위밋(We-meet) 프로젝트 참여 학생, [2]에듀템 대표이사, [3*]광운대학교 소프트웨어사업단 초빙교수

Deep Learning-based Virtual Tutor Program to Improve English Learning Skills for Beginners in English

Jae-Seong Ju[1] · Woo-Jin Hong[1] · Woo-Hyeok Moon[1] · Sa-Ra Yu[1] · Kyeong-Soo Ko[1] · Yeong-Seo Go[1] · Ha-Eun Jang[1] · Si-Hyeon Kim[1] · Jeong-Hoon Shin[2] · Seoung-Ho Choi[3*]

[1]We-meet project participating students
[2]CEO, Edutem
[3*]Visting Professor, National Program for Excellence in Software, Kwangwoon University, Seoul 018967, Korea

[요 약]

영어를 처음 배울 때, 발음은 익숙해도 타인과 대화를 수행하는 것은 어렵다. 이는 혼자서 영어 회화 연습이 어렵기 때문이다. 이러한 문제를 해결하기 위해 딥러닝을 기반으로 한 가상 튜터 프로그램을 제안한다. 이 프로그램은 영어 초보자들이 스스로 학습하는 상황에서 대화를 유도하기 위해 13가지 시나리오를 사용한다. 이 논문에서는 이러한 13가지 시나리오를 기반으로 약 3,200개의 영상을 직접 제작, 수집했다. 제안 프로그램은 먼저 딥페이크 기술을 사용하여 가상 인간의 표정과 발화 상황을 인식하고 생성한다. 그리고 립마우스 모델을 적용하여 정확한 발화 형태를 실현하여 영어 발화 교육을 보조한다. 이를 통해 생성된 가상 인간이 실제 사람의 발음과 유사하게 생성되었음을 확인할 수 있었다. 또한, 입 모양 모델을 적용한 가상 인간의 발화 정확도가 딥페이크 모델만 적용한 것보다 높음을 확인했다. 이를 통해, 실제 영어 초보자들에게 어려운 발음과 대화 문제를 경감시킬 수 있을 것이다.

[Abstract]

In learning English, it can be challenging to have conversations with others, even for those who are familiar with the pronunciation. Practicing communication with a real person is not easy for those who are just starting out. To address this issue, we propose a deep learning-based virtual tutor program for English beginners to practice speaking English on their own. The program uses 13 scenarios to simulate conversations, producing direct data. We used deepfake technology to create a virtual human who accurately mimics speech and facial expressions. We also applied a lip mouse model to enhance speech accuracy. Our experimental results showed that the virtual human's pronunciation closely resembles that of a real person, and the lip mouse model improves speech accuracy compared to when the deepfake model is used alone. Therefore, English beginners can improve their pronunciation and conversation skills in English using our program.

색인어 : 영어 초보자 학습 능력을 위한 교육 프로그램, 가상 인간, 딥 페이크, 립 마우스, 가상 튜터

Keyword : Educational Program for English Beginner Learning Skills, Virtual Human, Deepfake Technology, Lip Mouse Model, Artificial Intelligence Tutor

http://dx.doi.org/10.9728/dcs.2023.24.5.999

Received 05 February 2023; **Revised** 28 February 2023
Accepted 03 April 2023

***Corresponding Author; Seoung-Ho Choi**

Tel: +※ 개인정보 표시제한
E-mail: jcn99250@naver.com

Ⅰ. 서 론

영어는 전 세계의 공용어로 한국에서는 정규 교육인 초등부터 영어 읽기, 쓰기, 듣기, 말하기를 시작으로 다양한 문화적 배경의 사람들과 소통하는 능력을 키운다. 하지만 공교육에서의 영어 수업 내 말하기·쓰기 수업 비율은 평균 34.5%로 매우 낮게 나타났다. 이는 정량적 평가가 어려워 시험으로 시행되지 못하는 점, 단일민족의 비영어권 국가로 오르지 한국의 정규 교육과정에서 실제 원어민과 대화하는 경험이 턱없이 부족한 점을 꼽는다. 이에 2000년대 이후 학교 현장에 원어민 교사가 배치되는 원어민 영어 보조교사 초빙 프로그램(EPIK, English Program in Korea)이 시행되는 노력이 가속화되었다. 하지만 실제로는 비용, 내신 위주의 교육과정, 강사의 자질 및 도덕성 문제 등의 이유로 원활한 운영이 이뤄지지 않고 점차 원어민 강사 고용이 줄고 있다[1].

하지만 2015년 개정 영어과 교육과정에서도 '영어 의사소통 역량'은 글로벌 시민의식, 자기관리 같은 핵심역량 중 최우선으로 여겨지고 있으며 정규 교육이 끝난 뒤 사회에서도 국제화에 발맞춰 언어 구사 능력을 갖춘 인재가 환영받는다[2]. 이처럼 영어를 학문적 탐구 대상에서 벗어나 실제 사람과 대화가 가능한 의사소통의 매개체로 바라보는 움직임이 계속되고 있다. 또한, 코로나 19 바이러스로 비대면 의사소통의 확대, IT 기술의 발전으로 이제는 사람과의 소통을 넘어 가상인간을 활용해 디지털 학습 생태계가 구축되고 있다. 최근 인공지능을 중심으로 IT 기술을 도입한 영어 회화 교육 서비스에는 스피쿠스(Spicus), Ai 튜터(Ai Tutor), 스픽나우(SpeakNow) 등이 있다. 스피쿠스는 동영상으로 학습하고 인공지능 기반으로 레벨 테스트와 말하기 연습을 한 뒤 실제 튜터와 화상 회화를 하는 방식이고, AI 튜터의 경우에는 상황에 따른 시나리오를 기반으로 인공지능과 대화를 나눈다. 대화가 끝나면 말했던 문장을 바탕으로 발음, 표현 등을 알려주며 학습해나가는 형식이다. 스픽나우는 인공지능과 상황에 따른 자유 대화를 진행 하며 실시간으로 인공지능이 개선점을 제안하고, 대화가 끝나면 개선점을 알려주고 복습하는 형식으로 학습한다. 스픽나우의 '북리딩' 서비스에서는 실제 원어민, 인플루언서 등을 기반으로 만들어진 가상 인간이 책을 읽어주는 서비스 또한 제공한다. 이런 인공지능을 활용한 영어 교육 프로그램은 시공간에 제약받지 않고, 배포가 쉽다는 장점이 있다. 또한, 실제 사람과 대화하는 것이 아니므로 딱딱하다는 의견도 있지만, 반복 학습하기에 좋고, 실수와 틀리는 것에 대한 두려움이 덜하다는 의견이 있다.

본 연구에서는 딥러닝 기반의 딥페이크와 립싱크를 이용하여 정확한 영어 발음의 발화에 초점을 둔 AI 가상 튜터를 개발하는 것을 목표로 한다. 영어 말하기 학습에 도움을 주어 딥페이크 기술이 교육적 목적으로 활용할 수 있는 사례를 제시하고자 한다. 딥페이크 및 립싱크를 적용한 결과에 성능 평가를 진행하고 성능 향상을 위한 데이터 수집, 모델 적용, 전처리 등의 방법론을 탐색하는 것에 중점을 두었다. 그리고 영상 데이터를 직접 제작, 수집하여 추가적인 개발에 활용될 수 있는 것을 목표로 했다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로 가상인간, 딥페이크, 립싱크 기술의 연구 동향에 대해 살펴보고, 3장에서는 본 논문의 제안 방법으로 제안 가상 튜터를 활용한 영어 학습 과정을 소개한다. 4장에서는 프로그램 제작 방법을 소개하고, 5장에서는 각 기술을 적용했을 때 실제 인간과 유사한 정도를 비교 분석한다. 마지막으로 5장에서는 결론 및 활용 방안을 제시하며 결론을 맺는다.

Ⅱ. 관련 연구

최근 가상 인간(Virtual Human)은 컴퓨터 그래픽(Computer Graphics, CG)의 발전으로 인간과 유사하게 보이고 행동하는 컴퓨터 생성 3차원 인공지능 캐릭터이다. 특히 코로나바이러스 감염증-19(COVID-19)의 확산으로 사람들은 집에 있는 시간이 증가함에 따라 온라인 생활이 보편화되었고 이는 메타버스 산업의 성장으로 이어졌다. 최근 가상 인간은 단순히 인간 묘사에 그치지 않고 딥러닝의 발전, 이미지 처리 능력 향상, 촬영 기술 발달 등으로 인간의 표정, 제스처, 발화와 같은 멀티모달(Multi-modal)을 활용하여 인간을 표방하려는 노력이 이뤄지고 있다. 두 개의 카메라를 동시에 사용하여 사람의 자세와 표현을 동시에 탐지하여 3D 가상 인간 애니메이션을 자연스럽게 생성하는 방식을 제안했다[3]. 기존의 3D 신체 모델은 얼굴 표현이 단순하고 3D 얼굴 모델은 얼굴에 알맞은 신체를 합성하지 못한다는 단점을 극복하기 위해서 얼굴의 형태와 신체를 같이 모델링 하는 방법론을 제안했다[4]. 가상 인간 모델을 생성하는 과정에서 많은 시간과 비용이 발생하는 한계를 극복해 빠르고 낮은 성능의 환경에서도 사용할 수 있는 MDD-NET 기반의 가상 인간 모델 생성 기술을 제안했다[5]. 인공지능을 이용한 이미지 합성, 음성 합성 기술이 발전함에 따라 다양한 분야에서 가상 인간 모델을 활용하고 있다. 대표적으로 국내 기업 딥브레인 AI[6]는 인공지능 기술을 기반으로 텍스트를 입력하면 가상인간 동영상을 생성하는 기술, 다양한 상황에 적합한 대화형 AI 휴먼, AI 아바타와 같은 가상 인간 생성 솔루션을 제공한다. 이러한 서비스는 다양한 곳에서 사용되고 있는데 특히 교육 측면에서 실제 적용되고 있으며 관련된 연구도 활발하게 이루어지고 있다. 가상 현실 기술이 영어 교육에 적용되는 사례를 공교육 환경, 사교육 환경, 영어 교육 게임, 전문분야의 영어 교육 등의 상황으로 나누어 분석했다[7].

가상현실(Virtual Reality)은 지식을 획득하거나 학습하는 데 매우 효과적인 기술로 학습자가 시간과 공간을 초월하여 학습할 수 있는 매우 효과적인 학습 수단이다. 가상 튜터(Virtual Tutor)는 가상현실에서 사용되는 교육 전문가로 교수자를 대신하여 학습에 필요한 지식, 기능, 태도를 제공하여 학습 진행자의 임무를 수행한다. 가상 튜터를 활용한 학습 결

과는 학습자의 인지 부하를 감소하고 학습에 대한 동기와 지식에 관한 관심이 높아짐을 확인할 수 있다[8],[9],[10]. 이는 가상 튜터를 활용한 방법이 학습자에게 도움이 됨을 시사한다.

딥페이크(Deepfake)란, 인공지능 기술인 딥러닝과 가짜를 의미하는 단어인 페이크의 합성어이다. 본격적으로 해당 용어가 알려지게 된 사건은 2017년 reddit의 'deepfakes'라는 유저가 딥러닝 기술을 이용해서 유명 연예인의 얼굴을 음란 영상에 합성할 수 있다고 주장하는 글이 올라와 관심을 받으며 알려졌다. 딥페이크 콘텐츠는 최근 DeepFaceLab[11], FaceSwap[12]등과 같은 오픈 소스 형태의 영상 합성 제작 프로그램이 배포되면서 더욱 확산되고 있다. 또한 딥페이크 기술이 영상 산업에 적극 활용되고 있는데, AR/VR 영상을 제작하거나 간단하게 특수 효과를 만들어 내는 것이 그 예이다. 하지만 한편으로는 가짜 뉴스나 허위 음란물을 제작하는 등 기술의 악용으로 혼란을 일으키고 있기도 하다. 딥페이크는 영상 속 A의 얼굴을 B로 바꾸는 과정은 크게 추출(Extraction)-학습(Learning)-병합(Merging)의 세 단계로 진행된다. 여기서 합성하는 사람을 SRC, 합성 당하는 객체를 DST라 칭한다. 추출 단계는 기계가 얼굴을 탐지하는 과정으로 눈, 코, 입 등 특징점(랜드마크)을 탐지한다. 학습 단계는 추출 단계에서 추출한 사람의 얼굴 특징을 다른 얼굴에 재생성하는 과정이다. 해당 단계에서는 생성적 적대 신경망(Generatvive Adversarial Networks, GAN)을 이용하여 이미지의 얼굴 위에 다른 얼굴을 결합해 새로운 이미지를 생성하는 기술이다. GAN은 원본과 차이가 없는 데이터를 만들어내려는 생성자(Generator)와 원본과 다른 점을 찾아내는 판별자(Discriminator)를 서로 경쟁시켜 반복 실행을 통해 원본과 흡사한 데이터를 만들어내는 원리를 가진 기술이다. 실제 데이터를 학습하는 과정에서 생성자가 진짜와 유사한 가짜 데이터를 만들어내면 판별자가 해당 데이터의 진위를 판별하는 생성 알고리즘이다. 마지막 딥페이크의 마지막 단계는 병합으로 새롭게 만들어진 SRC의 얼굴을 DST에 합친다. 이때 추가적인 기계 학습은 이뤄지지 않고 후보정만 이뤄지게 된다. 최종적으로 DST 이미지의 얼굴 위에 SRC 이미지의 얼굴을 결합해 새로운 이미지를 생성한다.

딥페이크 기술에는 얼굴 합성(Face Synthesis), 속성 조작(Attribute Manipulation), 표정/표현 바꾸기(Expression Swap), 얼굴 바꾸기(Face Swap) 이 4가지 기술이 있다[13]. 전체 얼굴 합성은 존재하지 않는 얼굴 이미지 전체를 새로 만들어내는 기술이다. Karras 등은 AI가 완전히 새로운 가상의 영상을 생성하고, 표정이나 입모양 등 심층적인 조작을 가능하게 해주는 StyleGAN을 제안했다[14]. 두 번째, 표정/표현 바꾸기 기술은 얼굴의 표정을 임의로 조정하는 기술로 Thies 등은 RGB 데이터를 활용해 실시간(Real-time)으로 SOTA와 비슷한 성능의 모델을 생성하는 Face2Face를 제안했다[15]. 특징 바꾸기는 두 사람의 얼굴을 서로 바꾸는 방법으로 Perov 등은 기존 결과물의 성능이 자연스럽지 못한 경우가 많고 프레임워크의 처리 과정이 모호하거나 다양성이 부족하여 다양한 양의 데이터 셋에 따른 지원이 부족하다는 문제점을 보완하기 위해 명확한 파이프 라인을 갖춘 프레임워크인 DeepFaceLab을 개발하여 제안했다[11].

데이터 과학에서 립싱크(LipSync)란, 오디오와 비디오를 동기화하는 과정으로, 비디오 속 인물 혹은 캐릭터의 입술 움직임을 다른 오디오와 일치하는 과정이다. 립싱크는 시청각 콘텐츠가 기하급수적으로 증가함에 따라 빠른 영화 컨텐츠 제작이 중요해지며 영화, TV쇼, 비디오 게임 등 다양한 미디어 매체에서 시청자가 현실적이고 자연스럽게 영상을 보기 위해 이용된다. 립싱크는 얼굴 인식, 음성 인식, 감정 인식 및 자연어 처리와 같은 다양한 방면에서 혼재되어 사용된다. 통상적으로 사용되는 기계 번역을 위한 다른 언어로의 변환뿐만 아니라 가상현실에서 사실적인 3D 아바타를 제작하여 더욱 자연스럽게 상호작용할 수 있고, 음성 인식 시스템의 정확도를 개선하여 음성 명령을 더욱 잘 이해하고 응답하는 것에 도움을 줄 수 있다. 이는 음성과 입 모양만을 이용하여 학습하기에 언어뿐만 아니라 영상 처리를 통한 얼굴 및 목소리에도 입 모양을 동기화시킬 수 있다는 장점이 있다.

선행 연구에서는 입술 영역 내의 여러 정보를 특징으로 생성하는 방법, 입 모양을 근사화한 좌표를 기반으로 입술 움직임을 나타내는 벡터들의 집합을 특징으로 생성하는 방법, 상위의 두 가지를 결합한 방법 등 여러 가지의 방법을 통해 입 모양을 인식하고는 한다. 이를 위해서는 다음과 같은 단계를 거치고는 하는데, 보통 [16],[17] 크게 전처리, 얼굴 영역 및 발화 구간 검출과 같은 단계를 거쳐 생성된 이미지와 음성 데이터를 특징점을 통해 립싱크 모델을 작성하고는 한다.

Ⅲ. 제안 방법

그림 1은 최종 가상 튜터를 활용한 영어 학습 과정이다. 딥페이크와 립마우스 기술을 이용해 생성한 가상인간을 바탕으로 영어 초보자의 말하기 학습이 이루어진다. 영어 학습 단계는 4단계로 구성되었다. (1단계: 가상인간 발화, 2단계: 학습자 발화, 3단계: 학습자 발화 평가 4단계: 학습자 발화 피드백).

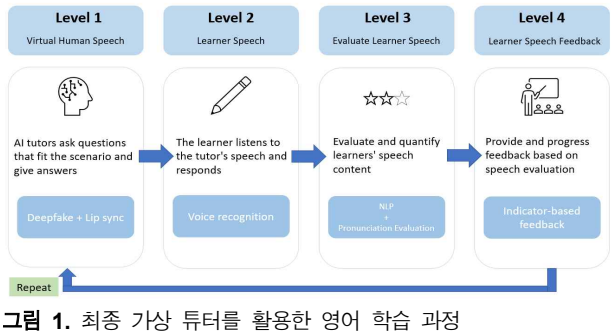

그림 1. 최종 가상 튜터를 활용한 영어 학습 과정
Fig. 1. Deep learning-based virtual tutor program system configuration

1단계는 가상인간 발화 단계로 가상인간이 학습자에게 다

양한 시나리오에 맞는 질문과 대답을 학습자에게 전달한다. 학습자는 가상인간의 발화를 듣고 올바른 발음에 대한 정보를 획득하고 가상인간의 발화를 바탕으로 발화를 준비한다.

2단계는 학습자 발화 단계로 학습자가 가상인간의 발화를 듣고 응답하는 단계이다. 가상인간이 발화한 내용을 바탕으로 학습자가 발화한다. 튜터 프로그램은 음성 인식 기술을 이용하여 학습자의 발화를 기록하고 다음 단계에서 평가에 사용할 수 있도록 준비한다.

3단계는 학습자 발화 평가 단계로 앞서 학습자가 발화한 내용을 평가한다. 앞서 기록된 학습자의 발화에 대하여 발음 및 내용 측면에서 평가를 진행한다. 앞서 기록된 음성에 대하여 자연어 처리와 발음평가를 위한 기술을 적용하여 학습자에게 피드백을 제공하기 위한 평가자료를 준비한다.

4단계는 학습자 발화 피드백 단계로 앞서 학습자의 발화에 대한 평가자료를 바탕으로 학습자에게 피드백을 제공한다. 발화의 빠르기, 발음의 정확도, 내용의 적절성 측면에 대해서 별점 혹은 점수의 형태로 지표를 제공하고 지표가 낮은 경우에는 학습자에게 다시 발화를 연습하도록 피드백을 진행한다. 일정 수준 이상으로 좋은 지표를 획득한 경우에는 새로운 학습으로 넘어가 다시 1단계부터 학습을 시작하게 된다.

해당 영어 학습 과정을 구현하기 위해서 본 논문에서는 원어민 튜터와 학습자가 대화 시에 발생하는 시간적, 공간적 제약을 줄임과 동시에 실제 인간과 대화하는 데 있어 학습자가 느끼는 부담을 완화하기 위하여 가상 인간 기반의 회화 튜터를 제안한다.

그림 2는 딥러닝 기반 가상 튜터를 생성하기 위한 플로우 차트이다. 첫째, 딥페이크 기술[12]을 통해 가상 AI의 외관을 생성한다. SRC, DST로 사용될 두 개의 영상 데이터를 생성한 이후 이미지를 추출하여 딥페이크 모델에 학습시킨다. 학습된 모델을 이용해 가상 인간 AI의 외관을 완성한다. 둘째, TTS(Text to Speech) 기술을 이용해 영어 대본을 발화 음성 파일의 형태로 가공하여 가상 인간의 목소리 데이터를 준비한다. 셋째, 립싱크 기술을 통해 영어 대본 발화 음성으로부터 AI의 입 모양을 생성한 뒤, 딥페이크를 통해 완성된 AI의 외관과 합성한다[12]. 이후 준비된 음성 데이터와 가상 인간 AI를 병합하여 대본의 내용을 발화할 수 있는 가상 튜터를 완성한다. 넷째, 대화 단계에 맞추어 사용자와 상호작용할 수 있으며, 사용자의 대답 음성을 녹화하여 음성 인식 기술을 통해 대본과 비교하여 피드백을 진행한다[18].

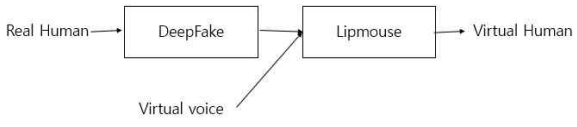

그림 2. 딥러닝 기반 가상 튜터 프로그램 시스템 구성도
Fig. 2. Deep learning-based virtual tutor program system configuration

IV. 실험 방법

시나리오는 사용자가 주제별로 학습할 수 있도록 가상 인간과의 대화 상황을 구상한 총 13개의 발화문으로 (인사, 시험공부1, 시험공부2, 면접, 사랑, 전화, 비즈니스, 대화와 응답, 의견파악, 관계, 할 일 · 부탁한 일, 숫자 정보1, 숫자 정보2) 이루어져 있다. 그림 3는 시험공부 1에 대한 시나리오 예시이다. 한 발화문은 Title - Description - Guide - Task(Reaction, Question, Guide, Answer) - Mission complete Reaction 구성으로 Task는 대부분 4개이지만 발화문에 따라 그 개수가 3개 혹은 5개로 다르다.

사용자와 가상 인간이 번갈아 가며 대화하는 형식의 프로그램이므로 가상 인간이 말하는 Title, Description, Reaction, Question, Mission complete Reaction 에 대해서 각 시나리오를 구성하였다. 총 13개의 발화문에 대하여, 가상 인간을 통해 전달되는 문장을 기반으로 발화 영상 데이터를 수집하였다. 분장 개수에 따라 각 시나리오에서 9개 혹은 11개의 영상이 수집되었고 '인사' 발화문의 경우 각 문장별 25개씩 8명의 데이터를 수집하여 1,800개의 영상이, 나머지 12개의 발화문의 경우 각 문장 별 1개씩 10명의 데이터를 수집하여 1,400개의 영상이 수집되었다.

Id			면접01
Level			초급
Category			비즈니스
Mission_Start		Title	A job interview
		Description_Ko	구직자인 당신은 희망하던 회사에 면접을 보러 왔습니다. 긴장되지만 침착하게 대답해 봅시다!
		Description_En	As an applicant, you came to the company you wanted for an interview. Of course you're getting nervous, but let's answer calmly!
		Guide_Ko	어서오세요.
		Guide_En	Welcome, get in.
Task_1	Basic_Task	Reaction	We'd like to ask you a few questions.
		Q1	Did you already graduate? Or when are you graduating?
		Guide	올해 2월에 ABC 대학을 졸업할 예정이라고 말해보세요.
		Answer	I am graduating from ABC University this February.
Task_2	Basic_Task	Reaction	Hmm...
		Q2	What are you majoring in?
		Guide	컴퓨터 공학을 전공하고, 비즈니스를 부전공했다고 말해보세요.
		Answer	My major is Computer Engineering, and I have a minor in Business.
Task_3	Basic_Task	Reaction	That's good.
		Q3	What made you apply for Software Development Department?
		Guide	이 자리가 당신의 전공과 관련된 학업적 지식을 실전에 적용하기 가장 좋은 곳이라고 대답해 보세요.
		Answer	This position is the best place for me to put my academic knowledge regarding my major into real practice.
Task_4	Basic_Task	Reaction	I see.
		Q4	Can you perform overseas duties as well?
		Guide	학교의 경쟁력 있고 다양한 해외 프로그램 덕분에 해외에서 공부할 기회가 많았다고 답해보세요.
		Answer	I had many opportunities to study abroad thanks to my school's competitive and diverse overseas programs.
Mission_Complete		Reaction	Alright.

(*) Korean interpretation of scenario is needed
그림 3. 시험공부 1에 대한 시나리오 예시
Fig. 3. Deep learning-based virtual tutor program system configuration

그림 4. 실제 데이터셋 영상 촬영 모습
Fig. 4. Deep learning-based virtual tutor program system configuration

그림 4에서 보이듯이 영상은 데스크탑의 web-camera 또는 휴대폰 카메라를 활용하여 해상도 360(640, 360) 이상의 mp4 형식으로 촬영되었다. 최종적으로 13개의 발화문에 대하여 총 3,200개의 영상이 수집되었다. 수집된 데이터는 발화자에 따라 폴더에 구분되어 저장되었고, 폴더 명의 경우 11개의 문자로 이루어져 있다. 영상의 경우 순서대로 5자리 숫자로 나타내었다(i.e 00001, 00002,…, N). 전체 데이터 중 70%는 학습 데이터로, 30%는 검증 데이터로 사용하였다.

모델 학습을 위해 다음과 같은 전처리 과정을 거친다. 우선 FFmpeg 라이브러리를 활용하여 영상(.mp4)에서 오디오 파일(.wav)을 추출하여 해당 오디오 파일의 속성을 음성 채널은 모노(Mono, 단일 채널)로, 주파수는 16kHz로 변경하여 저장한다. 영상의 경우 Python의 cv2 라이브러리를 활용하여 FRS를 25프레임으로 변경한 후, 영상을 프레임에 맞춰 이미지(.jpg)로 나눴다. Python의 retina-face 라이브러리를 활용하여 분할된 이미지에서 얼굴 부분을 검출한 후, 해당 부분을 추출하였다. 추출된 이미지를 모델 학습에 사용하기 위해, 이미지의 크기가 224*224보다 크면 Python의 cv2 라이브러리를 사용하여 크기를 조정하였다.

본 논문에서 사용한 딥러닝 모델은 총 두 가지를 사용하여 검증을 진행했다. 딥페이크 부분에서는 DeepFaceLab 모델[12]을 사용했다. 그리고 립싱크 부분에서는 사전 학습된 Wav2Lip 모델[17]을 사용했다.

V. 실험 결과

표 1 그리고 표 2는 코사인 유사도를 사용하여 목적 영상과 딥페이크를 적용한 영상과의 소스 영상의 유사도를 비교했다. 비교를 위해 opencv 라이브러리를 이용해 영상의 구간을 10개로 나누어 구간별로 사진을 추출한뒤에 deepface 라이브러리를 이용해 src 사진과의 코사인 유사도를 비교하고 평균 및 표준편차를 구하였다. dst의 평균이 딥페이크의 평균보다 높음을 볼 수 있는데, 아직은 딥페이크의 성능이 원본을 따라가지 못함을 알 수 있다.

본 연구의 가상 튜터에게 립마우스를 적용하면 음성과 발화 일치율이 증가한다는 것을 검정 수행하기 위해 측정 모형의 적합성 (및 타당성)을 확인하였다. 본 단계에서는 원본, 딥페이크(이하 DFL)를 적용한 영상, 립마우스(이하 W2L)를 적용한 영상의 입모양 생성 및 발화 정확도를 알아보기 위해 13개 시나리오의 LSE-D, LSE-C 값 변화를 알아보았다. LSE-D(Lip-Sync Error-Distance)는 그 값이 작을수록 오디오와 영상 간 입 모양 생성이 높은 일치를 보이고, LSE-C (Lip-Sync Error-Confidence)는 클수록 오디오와 영상 속 발화하는 모습이 높은 일치를 보이는 결과를 보인다. LSE 값을 측정하기 위해 사전학습된 SyncNet 모델을 사용했다. 표 3, 표 4, 표 5, 표 6는 성별에 따른 각 영상 단계별 시나리오의 구체적인 LSE-C, LSE-D 값과 변화율이다.

표 1. 남자 대상 딥페이크 영상과 원본사진의 코사인 유사도 결과
Table 1. Results of cosine similarity between deepfake video and orginal video for male subjects

Man Scenario	Dst		Deepfake	
	Mean	SD	Mean	SD
greeting	0.50	0.09	0.37	0.14
exam01	0.53	0.13	0.27	0.13
exam02	0.52	0.14	0.3	0.15
interview	0.49	0.1	0.3	0.2
love	0.48	0.09	0.27	0.15
call	0.48	0.1	0.28	0.15
business	0.46	0.08	0.35	0.16
conversation	0.45	0.08	0.29	0.18
opinion	0.46	0.11	0.35	0.14
relationship	0.47	0.1	0.31	0.16
todo	0.46	0.12	0.39	0.13
number01	0.47	0.09	0.32	0.15
number02	0.51	0.14	0.44	0.17

표 2. 여자 대상 딥페이크 영상과 원본사진의 코사인 유사도 결과
Table 2. Results of cosine similarity between deepfake video and orginal video for woman subjects

Woman Scenario	Dst		Deepfake	
	Mean	SD	Mean	SD
greeting	0.46	0.14	0.26	0.21
exam01	0.46	0.1	0.38	0.15
exam02	0.54	0.17	0.33	0.25
interview	0.56	0.15	0.32	0.25
love	0.54	0.15	0.32	0.25
call	0.56	0.14	0.33	0.26
business	0.53	0.16	0.33	0.26
conversation	0.49	0.19	0.17	0.19
opinion	0.58	0.12	0.34	0.25
relationship	0.56	0.14	0.29	0.24
todo	0.56	0.15	0.34	0.25
number01	0.58	0.13	0.34	0.25
number02	0.57	0.15	0.28	0.25

남성의 경우, 원본과 DFL을 적용한 영상 간에는 LSE-D가 평균 4% 하락, LSE-C가 평균 28% 상승하였고, 원본과 W2L을 적용한 영상 간에는 LSE-D가 평균 1.4% 하락, LSE-C가 평균 136% 상승하는 것이 확인됐다. 그 결과 두 방법 모두 유의미한 영향을 끼치는 것을 알 수 있다. 여성의 경우, 원본과 DFL을 적용한 영상 간에는 LSE-D가 평균 7% 상승, LSE-C가 평균 7% 하락하였고, 원본과 W2L을 적용한 영상 간에는 LSE-D가 평균 1% 상승, LSE-C가 평균 13% 상승하는 것이 확인됐다. 극적인 예시로, '면접' 시나리오에서 원본 영상에 DFL을 적용했을 때 오히려 점수가 소폭 낮아졌으나(LSE-D 0.27 상승, LSE-C 0.102 하락) W2L을 적용했을 때 점수가 눈에 띄게 증가(LSE-D 2.898 하락, LSE-C 4.468 상승)한 것을 확인할 수 있다. 결국, 가상 인간의 정확한 문장 전달을 목표로 영어 교육 프로그램에서 Wav2Lip으로 인해 발화 정확도가 증가한 것은 본 연구의 중요성을 시사한다.

표 3. 남자 대상 전체 시나리오에 대한 LSE-C 결과
Table 3. LSE-C results for all scenarios for man

Man LSE-C	Original Confidence (1)	DFL Confidence (2)	Wav2Lip Confidence (3)	{(1)-(2)}/100	{(1)-(3)}/100
greeting	1.945	1.311	2.997	-0.006	0.011
exam01	1.061	0.568	1.035	-0.005	0.000
exam02	0.370	1.365	1.175	0.010	0.008
interview	0.960	0.858	5.428	-0.001	0.045
love	0.870	0.662	1.323	-0.002	0.005
call	0.866	0.613	2.242	-0.003	0.014
business	0.742	1.224	2.053	0.005	0.013
conversation	1.016	1.391	1.301	0.004	0.003
opinion	0.574	0.633	0.450	0.001	-0.001
relationship	0.847	1.537	1.489	0.007	0.006
todo	0.527	0.904	1.474	0.004	0.009
number01	0.910	0.781	1.445	-0.001	0.005
number02	0.782	0.713	3.389	-0.001	0.026

표 4. 남자 대상 전체 시나리오에 대한 LSE-D 결과
Table 4. LSE-D results for all scenarios for man

Man LSE-D	Original Confidence (1)	DFL Confidence (2)	Wav2Lip Confidence (3)	{(1)-(2)}/100	{(1)-(3)}/100
greeting	8.776	9.852	7.746	0.011	-0.010
exam01	10.920	10.449	11.935	-0.005	0.010
exam02	12.374	9.149	11.862	-0.032	-0.005
interview	10.808	11.078	7.910	0.003	-0.029
love	12.230	12.834	13.189	0.006	0.010
call	11.815	11.942	10.656	0.001	-0.012
business	11.225	7.601	12.551	-0.036	0.013
conversation	10.872	9.549	12.488	-0.013	0.016
opinion	11.934	13.487	12.677	0.016	0.007
relationship	11.215	9.686	12.955	-0.015	0.017
todo	10.686	12.141	11.200	0.015	0.005
number01	12.801	11.891	11.434	-0.009	-0.014
number02	12.525	12.153	9.355	-0.004	-0.032

표 5. 여자 대상 전체 시나리오에 대한 LSE-C 결과
Table 5. LSE-C results for all scenarios for woman

Woman LSE-C	Original Confidence (1)	DFL Confidence (2)	Wav2Lip Confidence (3)	{(1)-(2)}/100	{(1)-(3)}/100
greeting	2.007	1.937	3.791	-0.001	0.018
exam01	1.253	0.786	0.929	-0.005	-0.003
exam02	1.538	1.602	2.459	0.001	0.009
interview	1.387	1.605	2.649	0.002	0.013
love	0.814	0.904	0.830	0.001	0.000
call	1.170	1.107	1.537	-0.001	0.004
business	1.556	1.834	1.253	0.003	-0.003
conversation	1.017	1.075	1.377	0.001	0.004
opinion	1.961	0.920	0.979	-0.010	-0.010
relationship	0.922	0.948	1.053	0.000	0.001
todo	1.193	1.482	1.836	0.003	0.006
number01	3.372	2.461	0.984	-0.009	-0.024
number02	2.369	1.204	1.491	-0.012	-0.009

표 6. 여자 대상 전체 시나리오에 대한 LSE-D 결과
Table 6. LSE-D results for all scenarios for woman

Woman LSE-D	Original Confidence (1)	DFL Confidence (2)	Wav2Lip Confidence (3)	{(1)-(2)}/100	{(1)-(3)}/100
greeting	9.516	9.862	7.578	0.003	-0.019
exam01	11.929	12.047	12.586	0.001	0.007
exam02	11.797	11.950	10.331	0.002	-0.015
interview	11.000	11.026	9.614	0.000	-0.014
love	11.783	11.728	11.749	-0.001	0.000
call	12.005	12.124	10.859	0.001	-0.011
business	11.071	11.007	10.869	-0.001	-0.002
conversation	12.458	12.364	11.306	-0.001	-0.012
opinion	9.620	13.614	12.605	0.040	0.030
relationship	12.519	12.719	12.085	0.002	-0.004
todo	13.024	12.579	11.919	-0.004	-0.011
number01	10.163	12.059	13.584	0.019	0.034
number02	10.560	13.123	12.571	0.026	0.020

그림 5는 남자 및 여자의 원본 영상으로부터 딥페이크가 적용된 결과 그리고 립 마우스까지 적용된 결과를 보여주고 있다. 립 마우스를 적용했을 때가 실제 입이 움직이는 현상을 좀 더 잘 보여주고 있다. 이를 기반으로 그림 6는 실제 본 연구에서 생성된 가상 인간과 실제 대화를 하는 모습을 보여주고 있다. 최종 결과물로 딥페이크와 립싱크를 적용하여 몇 개의 시나리오에 대해 동영상을 만들었다. 우선 남자 데이터로 딥페이크와 립싱크를 적용했을 때 결과 링크이다. https://youtu.be/UliginIpFf8 다음으로 여자 데이터로 딥페이크에 립싱크를 적용했을 때 결과 링크이다. https://youtu.be/6RD22Rn1d4s 아래 표에서는 딥페이크만 적용한 결과물 동영상을 시나리오, 성별에 따라 구분했다.

a) b)

그림 5. 딥페이크 및 립싱크가 적용된 결과,
a) 딥페이크 적용, b) 딥페이크에 립싱크 적용
Fig. 5. Result with deepfake and lip sync applied,
a) apply deepfake, b) apply lip sync to deepfake

그림 6. 가상 튜터와 학습자가 서로 이야기를 하는 모습
Fig. 6. Talking to a virtual tutor

표 7. 딥페이크만 적용했을 때 결과 동영상 링크
Table 7. Result video link when deepfake only applied

Scenario	Gender	Url
relationship	man	https://youtu.be/MMz77stghJY
relationship	woman	https://youtu.be/XZdxBwiKsPQ
conversation	man	https://youtu.be/eamMUh9t-iA
conversation	woman	https://youtu.be/2HVOtfePWm0
interview	man	https://youtu.be/Da9xbusBSlU
interview	woman	https://youtu.be/cUG11l82_DI
business	man	https://youtu.be/hqXVXlnN7CQ
business	woman	https://youtu.be/rODfv_iGWCg
love	man	https://youtu.be/nMqSffQjN-8
love	woman	https://youtu.be/H5jh-ygZ0Bg
number01	man	https://youtu.be/eRebDRPo6P8
number01	woman	https://youtu.be/MnFlbfF5Yig
number02	man	https://youtu.be/nd1sJEFK32k
number02	woman	https://youtu.be/A5On5nERJ3w
exam01	man	https://youtu.be/sTIBhMD2eNY
exam01	woman	https://youtu.be/gsadA_WHt-M
exam02	man	https://youtu.be/WEYwVUc-iR8
exam02	woman	https://youtu.be/dju9Gd0ts30
opinion	man	https://youtu.be/kopiN7RK_0Q
call01	man	https://youtu.be/IrrcRHF0Rt8
call01	woman	https://youtu.be/ZnVeP0PWocg
todo	man	https://youtu.be/9xGW02c5U-Y
todo	woman	https://youtu.be/_uJyPZ-CwfM

VI. 결 론

영어 초보자가 혼자서 스스로 학습하는 상황에서 영어의 학습 능력을 형상화하기 위해 딥러닝 기반 가상 튜터 프로그램을 제안한다. 가상 튜터 프로그램은 13가지의 시나리오에 기반을 두어 대화를 유도했다. 본 논문에서는 13가지의 시나리오로 약 3200개의 영상을 직접 제작, 수집하였다. 영상을 제작, 수집하는 과정에서 딥페이크를 적용한 결과물의 수준을 높이기 위해서는 조명, 각도와 같은 촬영 환경의 통일이 매우 중요함을 알 수 있었다. 그리고 딥페이크의 적용할 때 비슷한 얼굴형과 피부색이 유사할수록 좋은 결과물을 얻을 수 있음을 확인했다. 제안 프로그램은 먼저 딥페이크 기술을 활용하여 가상 인간의 표정과 발화 상황을 인지하여 생성한 후 영어 발화 교육에 도움이 되도록 정확한 발화 모양 구현을 위해 립 마우스 모델을 적용했다.

추후 딥페이크 부분에서는 이마 부분까지 정확하게 딥페이크가 되는 방법을 고안하고자 하고 립 싱크 부분에서는 움직이는 중간에 끼지는 현상을 줄이는 방향으로 연구를 진행하고자 한다.

참고문헌

[1] S. Park and J. Kim, “Effects of English Lessons Taught in English: Pilot Study on Native and Korean English Teachers,” *The Journal of Mirae English Language and Literature*, Vol. 2, No. 27, pp. 101-118, May 2022. https://doi.org/10.46449/MJELL.2022.05.27.2.101

[2] Y. A. Lee, “Development of 2022 Revised English Curriculum: Issues and Challenges,” *Modern English Education*, Vol. 23, No. 1, pp. 28-41, February 2022.

[3] J. Wang, X. Chai, H. Guo, and J. Wang, “Generating 3D Virtual Human Animation Based on Facial Expression and Human Posture Captured by Dual Cameras,” *Advances in Multimedia*, 2022. http://lps3.doi.org.libproxy.snu.ac.kr/10.1155/2022/4833436

[4] N. Fang, L. Qiu, S. Zhang, Z. Wang, Y. Wang, Y. Gu, and J. Tan, “A Modeling Method for the Human Body Model with Facial Morphology,” *Computer-Aided Design*, Vol. 141, December 2021. https://doi.org/10.1016/j.cad.2021.103106

[5] N. Fang, L. Qiu, S. Zhang, Z. Wang, Y. Gu, and K. Hu. “The Rapid Construction Method of Human Body Model for Virtual Try-on on Mobile Terminal Based on MDD-Net,” *Soft Computing,* Vol. 26, pp. 12023-12039, August 2022. https://doi.org/10.1007/s00500-022-07464-3

[6] DEEPBRAIN AI. DEEPBRAIN AI Company Homepage [Internet]. Available: https://www.deepbrainai.io/

[7] Z. Pan, Y. Sun, Z. W. Yao, and M. Li, “Application of Virtual Reality in English Teaching,” in *Proceedings of 2021 3rd World Symposium on Artificial Intelligence (WSAI),* Guangzhou, China, pp. 64-71, June 2021. https://doi.org/10.1109/WSAI51899.2021.9486322

[8] A. L. Baylor and Y. Kim, “Simulating Instructional Roles Through Pedagogical Agents,” *International Journal of Artificial Intelligence in Education*, Vol. 15, No. 2, pp. 95-115, June 2005.

[9] S. Dinçer and A. Doğanay, “The Effects of Multiple-pedagogical Agents on Learners’ Academic Success, Motivation, and Cognitive Load,” *Computers & Education*, Vol. 111, pp. 74-100, August 2017. https://doi.org/10.1016/ j.compedu.2017.04.005

[10] S. Park, “The Effects of Social Cue Principles on Cognitive Load, Situational Interest, Motivation, and Achievement in Pedagogical Agent Multimedia Learning,” *Journal of Educational Technology & Society*, Vol. 18, No. 4, pp. 211-229, October 2015.

[11] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, ... and W. Zhang, “DeepFaceLab: Integrated, Flexible and Extensible Face-swapping Framework,” *arXiv Preprint arXiv:2005.05535*, May 2020. https://doi.org/10.48550/ arXiv.2005.05535

[12] S. Tariq, S. Lee, and S. Woo, “One Detector to Rule Them All: Towards a General Deepfake Attack Detection Framework,” in *Proceedings of the Web Conference 2021*, pp. 3625-3637, April 2021. https://doi.org/10.1145/3442381.3449809

[13] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection,” *Information Fusion,* Vol. 64, pp. 131-148, June 2020. https://doi.org/10.48550/arXiv.2001.00179

[14] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and Improving the Image Quality of StyleGAN,” in *Proceedings of Conference on Computer Vision and Pattern Recognition*, pp. 8110-8119, March 2020. https://doi.org/10.48550/arXiv.1912.04958

[15] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time Face Capture and Reenactment of RGB Videos,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Las Vegas, Nevada, pp. 2387-2395, June 2016. https://doi.org/10.48550/arXiv.2007.14808

[16] Y. K. Kim, J. G. Lim, and M. H. Kim, “Lip Reading Method Using CNN for Utterance Period Detection,” *Journal of Digital Convergence,* Vol. 14, No. 8, pp. 233-243, August 2016. https://doi.org/10.14400/JDC.2016.14.8.233

[17] K. R. Prajwal, and R. Mukhopadhyay, V. Namboodiri, C. V. Jawahar, “A Lip Sync Expert Is All You Need for Speech To Lip Generation in the Wild,” in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, pp. 484-492, October 2020. https://doi.org/10.48550/arXiv.2008.10010

[18] L. Muda, M. Begam, and I. Elamvazuthi, “Voice Recognition Algorithms Using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques,” *Journal of Computing*, Vol. 2, No. 3, March 2010. https://doi.org/10.48550/arXiv.1003.4083

주재성(Jae-Seong Ju)

2022년 : 서울대학교 전기정보공학부 학사과정

2023년~현 재: 서울대학교 ai연구원 인턴 연구원
※관심분야 : 딥러닝(Deep Learning), 강화학습(Reinforcement Learning), 데이터 시각화(Data visualization) 등

고경수(Kyeong-Soo Ko)

2022년 : 전북대학교 통계학과 학사과정

※관심분야 : 컴퓨터 비전(Computer Vision), 딥러닝(Deep Learning), 데이터 사이언스(Data Science) 등

홍우진(Woo-Jin Hong)

2022년 : 경기과학기술대학교 인공지능학과 학사과정

※관심분야 : 컴퓨터 비전(Computer Vision), 자연어 처리(Natural Language Processing), 강화학습(Reinforcement learning) 등

고영서(Yeong-Seo Go)

2022년 : 한동대학교 AI·컴퓨터공학심화전공 학사과정

※관심분야 : 빅데이터(Big Data), 데이터 사이언스(Data Science), 데이터 엔지니어링(Data Engineering) 등

문우혁(Woo-Hyeok Moon)

2022년 : 전북대학교 통계학과 학사과정

※관심분야 : 데이터 사이언스(Data Science), 응용통계(Applied Statistics), 딥러닝(Deep Learning) 등

장하은(Ha-Eun Jang)

2022년 : 한동대학교 커뮤니케이션 학부 학사과정

2018년~현 재: 한동대학교 커뮤니케이션학부 학사과정
※관심분야 : 컴퓨터 비전(Computer Vision), 딥러닝(Deep Learning), 데이터 사이언스(Data Science) 등

유사라(Sa-Ra Yu)

2022년 : 숙명여자대학교 컴퓨터과학과

2021년~2023년: 숙명여자대학교 데이터 지능 연구실 인턴 연구원
2023년~현 재: 숙명여자대학교 컴퓨터과학과 석사과정
※관심분야 : 데이터 마이닝(Data Mining), 데이터베이스(DataBase), 딥러닝(Deep Learning) 등

김시현(Si-Hyeon Kim)

2022년 : 서울시립대학교 도시행정학과 학사과정

2020년~현 재: 서울시립대학교 도시행정학과 학사 과정
※관심분야 : 딥러닝(Deep Learning), 강화학습(Reinforcement Learning), 자연어 처리(Natural Language Processing) 등

신정훈(Jeonghun Shin)

2016년～현　재: 에듀템 대표이사

최승호(Seoung-Ho Choi)

2018년 : 한성대학교 전자정보공학과 (공학사)
2020년 : 한성대학교 전자정보공학과 (공학석사)

2021년～현　재: 한성대학교 기초교양학부 시간강사
2022년～현　재: 광운대학교 소프트웨어 사업단 초빙교수