한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 28, No. 3, P. 23~29

감정에 기반한 가상인간의 대화 및 표정 실시간 생성 시스템 구현

김기락[1O] 연희연[2] 은태영[3] 정문열[1*]

[1]서강대학교 아트&테크놀로지, [2]인공지능학과, [3]컴퓨터공학과

{kirak, yeeun214, tyeun7, moon}@sogang.ac.kr

Emotion-based Real-time Facial Expression Matching Dialogue System for Virtual Human

Kirak Kim[1O] Heeyeon Yeon[2] Taeyoung Eun[3] Moonryul Jung[1*]

[1]Sogang University of Art&Technology, [2]Artificial Intelligence, [3]Computer Science Dept

요 약

가상인간은 가상공간(가상 현실, 혼합 현실, 메타버스 등)에서 Unity와 같은 3D Engine 전용 모델링 도구로 구현된다. 실제 사람과 유사한 외모, 목소리, 표정이나 행동 등을 구현하기 위해 다양한 가상인간 모델링 도구가 도입되었고, 어느 정도 수준까지 인간과 의사소통이 가능한 가상인간을 구현할 수 있게 되었다. 하지만, 지금까지의 가상인간 의사소통 방식은 대부분 텍스트 혹은 스피치만을 사용하는 단일모달에 머물러 있다. 최근 AI 기술이 발전함에 따라 가상인간의 의사소통 방식은 과거 기계 중심의 텍스트 기반 시스템에서 인간 중심의 자연스러운 멀티모달 의사소통 방식으로 변화할 수 있게 되었다. 본 논문에서는 다양한 대화 데이터셋으로 미세조정한 인공신경망을 사용해 사용자와 자연스럽게 대화 할 수 있는 가상인간을 구현하고, 해당 가상인간이 생성하는 문장의 감정값을 분석하여 이에 맞는 표정을 발화 중에 나타내는 시스템을 구현하여 사용자와 가상인간 간의 실시간 멀티모달 대화가 가능하게 하였다.

Abstract

Virtual humans are implemented with dedicated modeling tools like Unity 3D Engine in virtual space (virtual reality, mixed reality, metaverse, etc.). Various human modeling tools have been introduced to implement virtual human-like appearance, voice, expression, and behavior similar to real people, and virtual humans implemented via these tools can communicate with users to some extent. However, most of the virtual humans so far have stayed unimodal using only text or speech. As AI technologies advance, the outdated machine-centered dialogue system is now changing to a human-centered, natural multi-modal system. By using several pre-trained networks, we implemented an emotion-based multi-modal dialogue system, which generates human-like utterances and displays appropriate facial expressions in real-time.

키워드: 가상인간, 멀티모달 대화, 유니티, 감정기반 대화, GPT-2, RoBERTa

Keywords: Virtual Human, Multi-Modal Dialogue, Unity, Dialogue based on Emotions, GPT-2, RoBERTa

*corresponding Author: Moon-ryul Jung/Sogang University(moon@sogang.ac.kr)

*corresponding author: Moonryul Jung/Sogang University(moon@sogang.ac.kr)

Received : 2022.06.16./ Review completed : 1st 2022.06.29. / Accepted : 2022.07.06.
DOI : 10.15701/kcgs.2022.28.3.23
ISSN : 1975-7883(Print)/2383-529X(Online)

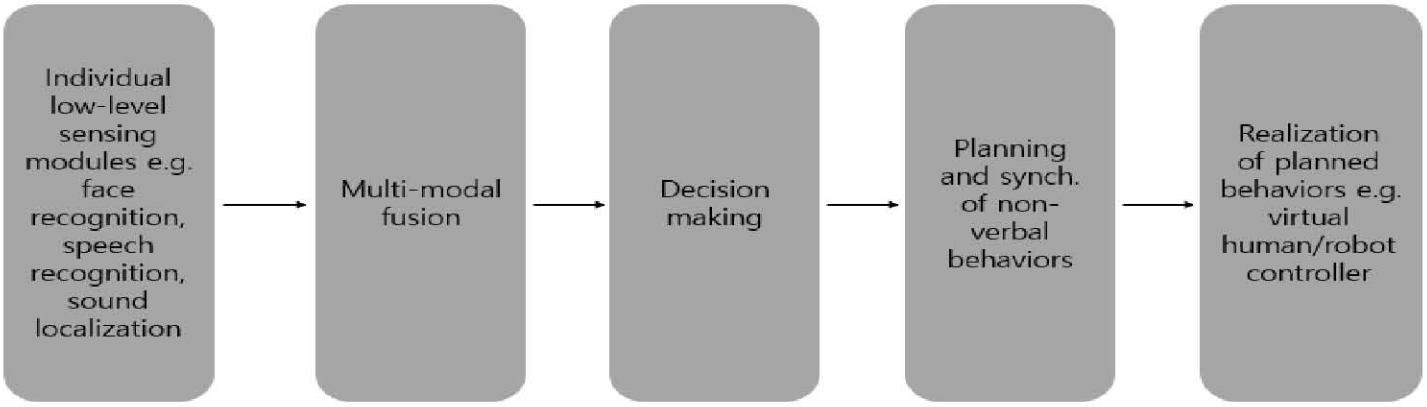

Figure 1 : Interaction with virtual humans : Overall steps

1. 서론

메타버스 시대가 도래함에 따라 멀티모달 대화 능력을 가진 가상인간의 중요성이 더욱 커지고 있다. 멀티모달 대화란 발화와 함께 표정, 제스처 등이 동반되어 다양한 감각양상이 사용되는 대화이다[1]. 현재 가상인간의 대화 시스템은 대부분 텍스트만을 사용자와 주고받는 단일모달 시스템이거나, 사전에 텍스트와 매칭된 제스처 및 표정을 재생하는 제한적인 방식으로 구현되어 있다. 사전에 매칭된 멀티모달 대화는 빠르게 변화하는 메타버스 환경에 대처할 수 없다. 따라서 실시간으로 발화에 맞는 제스처 및 표정을 생성하는 멀티모달 대화 시스템이 필요하다[2]. 또한 멀티모달 가상인간의 구성요소에는 컴퓨터 그래픽스, 애니메이션, AI, 인간과 컴퓨터의 상호작용 등 많은 요소들이 필요하다. IVA 2016 Tutorial[3]에서 가상인간이 갖춰야할 요구사항 중에는 5가지를 <Figure 1>에서 언급하고 있다. 본 논문에서는 해당 요구사항들 중 'Multi-modal fusion' 및 'Decision making'에 집중하여, 감정을 기반으로 발화에 맞는 얼굴 표정을 생성하는 멀티모달 대화 시스템을 개발하였다. 해당 시스템 개발을 위해 Unity 엔진을 사용하였다. Unity는 본래 게임 엔진으로 가상인간을 구현하기에 기본적인 컴포넌트가 많이 제공되고 있고 다른 게임/그래픽 엔진 OGRE3D나 Unreal Engine에 비해 멀티플랫폼(Mac OS, Xbox, PS4,5)을 지원하여 많은 곳에 배포할 수 있는 장점이 있다. 무엇보다도 Unity는 모듈형인 것이 큰 장점이다. 마치 레고 블록으로 조립하듯 사전에 구현된 부품들을(립싱크, TTS 등) 조립하여 가상인간을 제작할 수 있다. 이와 같이 전문가가 아니면 처음부터 구현하기 힘든 각종 물리 역학, 3D 모델링, 애니메이션과 같은 기능을 이미 있는 모듈들을 사용해 보다 쉽게 융합시킬 수 있는 장점을 지닌 툴이다.

본 논문에서는 Unity 3D Engine 상에서 C# 기반의 스크립트를 통해 UI 및 가상인간을 구현하고, 대화 및 감정값을 생성하는 Python 모델로부터 UDP Socket 통신을 통해 그 값을 받도록 하여 멀티모달 대화 시스템을 구현하였다. 대화 생성을 위해서 여러 일상대화 데이터셋들로 미세조정된 GPT-2 모델을 사용했고, 문장 감정값 분석을 위해서는 감정분류 데이터셋으로 미세조정된 RoBERTa 모델을 사용하였다.

본 논문의 공헌은 다음과 같다. 1)다양한 데이터셋으로 미세조정된 인공신경망과 Unity를 연동시켜 자연스러운 대화가 가능한 가상인간을 구현하였다. 2)미세조정된 RoBERTa와 EMFACS(EMotional Facial Action Coding System) 기반 Blendshape 모델을 연동시켜 가상인간이 발화에 맞는 표정을 실시간으로 짓게 하여 멀티모달 대화를 구현하였다.

2. 배경 지식 및 관련 연구

2.1 배경 지식

2.1.1 인공신경망

인공 신경망(ANN)[4]은 기계학습과 인지과학에서 생물학의 신경망(특히 뇌)에서 영감을 얻은 통계학적 학습 알고리즘이다. 인공신경망은 시냅스의 결합으로 네트워크를 형성한 인공 뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜, 문제 해결 능력을 가지는 모델 전반을 가리킨다(RNN, CNN 등).

2.1.2 Paul Ekman의 6가지 기본 감정

Paul Ekman 박사는 인간의 기본 감정을 6가지(분노, 놀람, 혐오감, 즐거움, 두려움, 슬픔)로<Figure 2> 구분했다. 이는 감정 관련 연구에 폭넓게 쓰이고 있으며, 추후 본문에서 언급될 얼굴 동작 코딩 시스템(FACS)의 기반이 되는 이론이다.

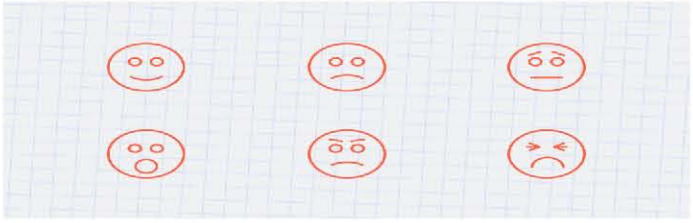

Figure 2 : Paul Ekman 6 emotions

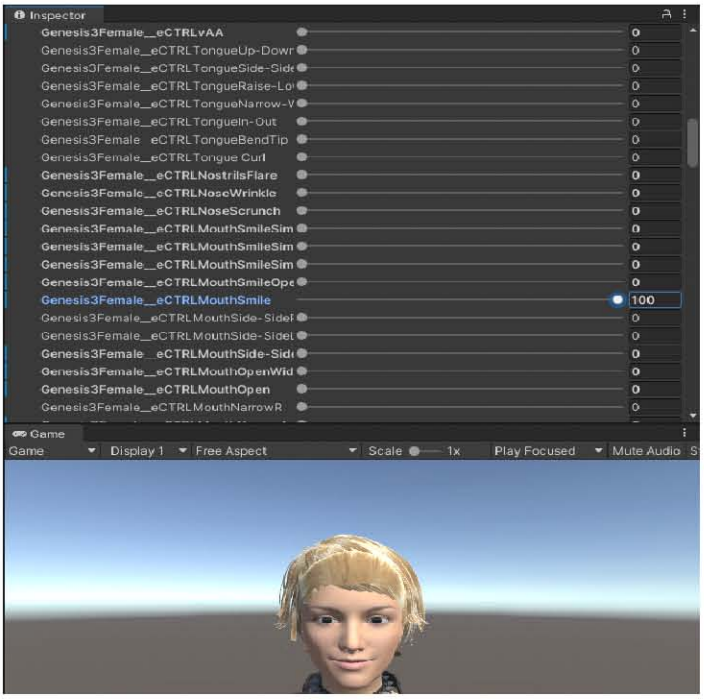

Figure 3: Blendshape targets on Unity

2.1.3 Blendshape

Blendshape은 얼굴 표정을 생성하는 선형 모델이며 주로 3D 캐릭터의 표정을 모델링하는데 사용된다[5]. Blendshape 모델은 해당 선형 모델의 basis vector에 해당하는 Blendshape target과 그들의 weight로 구성된다. 각 Blendshape target은 얼굴 Mesh에 일어나는 특정 변형(주로 의미적 변형을 뜻한다)을 의미하며, weight는 해당 변형이 얼마나 심하게 일어나는지를 0부터 1까지의 값으로 설정한다.

<Figure 3>을 보면 위의 Inspector에 현재 선택된 가상인간의 Blendshape target들이 나오며 이들의 weight 값을 슬라이더로 설정할 수 있다. Blendshape 모델을 수식으로 표시하면 다음과 같다 <Figure 4>. f는 임의의 표정, n는 해당 Blendshape 모델의 Blendshape target 개수, b_k 는 k번째 Blendshape target, 그리고 w_k는 그것의 weight으로 0부터 1까지의 값을 가진다.

$$f = \sum_{k=0}^{n} b_k w_k$$

Figure 4: Blendshape model equation

Blendshape target을 정의하는 방법은 다양하지만 주로 Facial Action Coding System(FACS)의 Action Unit에 기반하여 정의한다[6].

2.1.4 Facial Action Coding System(FACS)

FACS는 사람의 얼굴 표정을 분류하기 위해 고안된 체계이다[7]. FACS는 사람의 특정 표정을 해부학적으로 정의된 Action Unit(AU)들의 조합으로 설명한다. AU는 시각적으로 구별될 수 있는 가장 작은 얼굴 근육의 움직임이다[8]. EMFACS (EMotional Facial Action Coding System)은 감정에 관련된 얼굴 표정을 다루는 FACS이다[9]. EMFACS에서 설명되는 감정에 따른 facial action은 아래의 <Table 1>과 같다.

Table 1: Emotional action units

Emotion	Action Units
Happiness	6+ 12
Sadness	1+ 4+ 15
Surprise	1+ 2+ 5B+ 26
Fear	1+ 2+ 4+ 5+ 7+ 20+ 26
Anger	4+ 5+ 7+ 23
Disgust	9+ 15+ 17

2.2 관련 연구

2.2.1 GPT

대화 생성을 위해서는 GPT2 모델을 활용 했다. Generative Pre-Training(GPT-1)[10]은 대부분의 딥러닝 Task에서 발생하는 문제인 수동으로 라벨링된 많은 양의 데이터가 필요한 문제점의 한계에서 벗어나고자 비지도 사전 학습과 지도 미세조정을 결합한 준지도학습적 접근을 사용하였다. 약간의 조정만으로 다양한 종류의 task에서도 활용될 수 있는 범용 표현을 2단계를 거쳐 학습한다. 기존 RNN 등에 비해 구조화된 메모리를 쓸 수 있는 장점이 있다. 2018년에 최초 발표되어 현재는 인류 역사상 가장 뛰어난 인공지능이라고 평받는 GPT-3까지 발전했다.

2.2.2 RoBERTa

감정 분류를 위해서는 RoBERTa 모델을 사용했다. A Robustly Optimized BERT Pretraining Approach(이하 RoBERTa)에 기반이 되는 BERT는 구글에서 개발한 자연어 처리 신경망 구조이며 신경망을 통해 다음문장 예측(NSP)과 문장에서 가려진 단어(토큰)을 예측(MLM)할 수 있다. RoBERTa[11]는 기존 BERT 모델에서 Hyper Parameter 및 학습 데이터 사이즈 등을 조정함으로써 성능을 높인 모델이다. BERT는 과소적합 되는 경우가 많고 다양성에 취약하다. 또한 학습(training)과 추론(inference)이 분리되어 있기 때문에, 지속적 학습(continual learning)이 간단하지 않다. 따라서 RoBERTa는 기존의 BERT 모델에서 NSP-손실을 제거하고 훈련 데이터 셋의 fit을 조절하고, 긴 단어 추가 및 동적 마스킹(기존의 BERT에서는 매 학습 단계에서 똑같은 마스

크를 보게 되는데 RoBERTa에서는 매 epoch 마다 마스크를 새로 씌우는)을 적용하는 등의 여러 가지 미세조정을 거쳐서 성능을 높였다.

3. 모델 설명 및 구현 방법

3.1 가상인간 구현

본 논문에서는 가상인간 캐릭터 모델로 3D virtual character 제작에 특화된 툴인 Daz3d의 Genesis 3 female 모델을 사용하였다.

Genesis 3 female의 얼굴 Blendshape을 EMFACS(Emotional Facial Action Coding System)에 맞춰 weight값을 설정하여 Paul Ekman의 6가지 감정에 맞는 얼굴 표정을 <Figure 5>과 같이 정의하였다

TTS로는 스코틀랜드에 위치한 음성 합성 기술을 연구하는 회사인 Cereproc의 Cerevoice SDK 6.1.0을 사용하였다. Cerevoice는 고품질의 합성 음성을 제공할 뿐만 아니라 합성된 음성의 음소들과 그 타이밍도 제공하기 때문에 이를 이용하여 lip sync까지 구현할 수 있었다.

Figure 5: Implemented facial expressions for each emotions

3.2 모델 설명

3.2.1 대화생성

본 연구에서는 대화 생성을 위해 GPT-2을 사용하였다. GPT-2 모델은 자기지도학습 방식으로 매우 큰 영어 데이터 말뭉치에 대해 사전 훈련된 트랜스포머 모델이다. 학습 시에 문장에서 다음 단어를 추측하도록 훈련된 모델이다. 입력은 특정 길이의 연속 텍스트의 시퀀스이며 대상은 동일한 시퀀스로, 하나의 토큰(단어 또는 단어 조각)을 오른쪽으로 이동시킨다. 이 모델은 내부적으로 마스크 메커니즘을 사용하여 토큰 i에 대한 예측이 1부터 i까지의 입력만 사용하고 미래 토큰은 사용하지 않도록 한다.

이러한 방식으로 모델은 다운스트림 작업에 유용한 기능을 추출하는 데 사용할 수 있는 영어의 내부 표현을 학습한다. 그러나 이 모델은 사전 학습된 내용에 가장 적합하며, 프롬프트를 통해 텍스트를 생성하는 모델이다. 이러한 모델을 가상인간과의 chit chat 대화에 적합한 open domain 데이터셋에 대하여 미세조정을 진행하였다. 따라서 미세조정된 모델은 사용자의 발화에 대응하는 재치있는 답변을 생성할 수 있도록 적용된다. 대화생성에 학습한 데이터 셋은 1)DailyDialog[12], 2)EmpatheticDialogues[13], 3)Persona-Chat[14], 4)BlendedSkillTalk[15]을 사용했다.

3.2.2 감정분류

GPT-2로 부터 생성된 가상인간의 답변 텍스트를 Input로 받아 감정을 예측하는 모델로 RoBERTa 기반 모델의 distilled 버전 모델인 DistilRoBERTa 모델을 감정 분류 데이터셋을 활용해 추가 미세조정하여 활용했다. 모델은 6개의 레이어, 768차원 및 12개의 헤드를 가지고 있으며, 총 82M개의 매개 변수를 가지고 있다. 평균적으로 DistilRoBERTa는 Roberta 기반보다 두 배 더 빠르다. 본 모델을 감정 분류를 위해 감성 분류 데이터 셋에 대하여 미세조정을 진행하였고, Ekman의 6 감정에 기반한 7개의 감정 클래스(anger, disgust, fear, joy, neutral, sadness, surprise)로 분류하는 모델을 사용했다. 감정 분류 학습을 위해 감정 라벨이 달린 다양한 데이터 셋을 조합하여 감정 클래스별 불균형을 해소하며 학습하는 방식을 취했다. 미세조정에 사용된 데이터 셋은 1)Crowdflower(2016), 2)Emotion Dataset, Elvis et al.(2018), 3)GoEmotions, Demsky et al.(2020), 4)ISEAR, Vikash(2018), 5)MELD, Poria et al.(2019), 6)SemEval Mohammad et al.(2018)들의 조합이며, 본 데이터들은 트위터, Reddit, 보고서, TV 대사 등의 발언에서의 감정 레이블을 포함하고 있다. 하지만 모든 데이터가 동일하게 7개 감정 클래스를 포함하고 있는 것은 아니기 때문에 데이터 세트(감정당 2,811개의 관측치, 즉 총 20k에 가까운 관측치)에서 불균형성을 해소 후 학습을 진행했다. 공개된 평가 정확도는 66%이다.

4. 실험 방법 및 결과

20대 대학생 10명을 대상으로 실험이 진행되었다. 미세조정된 각 신경망들의(GPT2, RoBERTa) 효과를 검증하기 위해 2종류의 대조군을 만들어 총 3가지의 시나리오에서 실험하였다. 시나리오1은 미세조정 없이 GPT-2로만 대화한 것, 시나리오2는 미세조정한 GPT-2 그리고 시나리오3은 그것에 RoBERTa까지 적용한 것이다. 실험참여자로 하여금 자유롭게 본 논문의 가상인간과 자유롭게 대화를 하게 한 후 해당 경험에 대해 묻는 형태로 진행되었다. 해당 문항별 만족도

Figure 6: A conversation and matching facial expressions on each scenario

를 Likert Scale 5점 척도(1 : 매우 좋지 못함 - 5: 매우 만족함)에 따라 평가하였다.

제공된 질문 리스트는 <Table 2>와 같으며 만족도 설문 결과는 <Table 3>, <Figure 7>과 같다.

1번 시나리오는 미세조정 하지 않은, 입력에 대하여 다음에 나올 높은 확률의 단어가 단순하게 생성되는 형식의 GPT-2 모델과 감정 분류 모델 값을 활용하지 않은 시나리오이다. 따라서 가상인간은 맥락을 고려하지 못하고 대화의 연속성을 유지하지 못하였다. 실제로 평가 결과도 자연스럽지 못한 대화 상황에 불만족하는 경향을 보였다.

2번 시나리오는 미세조정한 GPT-2 모델을 사용했지만 감정 분류 모델 값을 활용하지 않은 시나리오이다. 본 시나리오에서는 가상인간은 맥락에 기반한 대화는 생성하지만 일관된 표정으로 부자연스러움을 연출하였다. 평가 결과 대화의 만족도 부분은 1번 시나리오에 비해 크게 향상되었지만 표정의 자연스러움이나 사람답지 못하다는 경향을 보였다.

3번 시나리오는 미세조정한 GPT-2 모델과 감정 분류 모델 값이 모두 활용된 시나리오이다. 본 시나리오에서 가상인간은 사용자의 발화에 맥락에 기반한 적절한 답변을 생성할 뿐만 아니라 그에 맞는 얼굴 표정이 드러냈다. 실제 평가결과 대화의 만족성 자체는 2번 시나리오와 비슷하나 훨씬 사람답다고 평가자들은 느꼈다.

사용자가 동일한 내용을 발화할 때의 각 시나리오별 대화 생성, 표정 매칭의 예시는 <Figure 7>과 같다.

5. 결론 및 향후 과제

5.1 결론(Conclusion) 및 평가(Discussion)

본 논문에서는 Unity, Cerevoice, Daz3D, GPT-2, RoBERTa 등을 이용하여 감정 기반 멀티모달 대화형 가상인간을 구현하였다. 학습 데이터 셋과 질문의 방향에 따라 사용자 경험의 만족도가 차이나지만, 실험을 통해 다양한 상황에서 사람과 같이 자연스런 대화가 가능하다는 것을 확인할 수 있었다.

Table 2: Likert scale question list

순번	질문
1	대화의 전반적 만족도(engaging)
2	대화 진행의 자연스러움(naturalness)
3	대화의 심도성(depth)
4	가상인간의 표정과 대화의 어울림
5	가상인간의 사람같은 정도 (친밀도/humanity)

Table 3: Total average user satisfaction on each scenario

1번 시나리오	2번 시나리오	3번 시나리오
1.52	2.88	3.84

Figure 7: Average user satisfaction per question on each Scenario

다만 현 시스템에서는 Unity와 Python의 정보 교환을 UDP 통신으로 하여 지연(Lag)이 발생하는 문제가 있다. 신경망에서 대답을 생성하는 시간도 존재하기 때문에 대화 사이에 공백이 어느 정도 있는 것이 자연스러워 사용자들이 실험 중에 통신으로 인한 지연으로 불편을 호소하지는 않았다. 다만 빠르게 환경이 변화하는 메타버스에서는 잠깐의 지연이 사용자 경험을 크게 저하할 수 있기 때문에 추후 연구에서는 인공신경망을 ONNX 파일로 변환하여 Unity에 임베딩하는 식으로 통신과정을 사라지게 해 지연 문제를 없앨 예정이다.

5.2 향후 과제

현재 시스템의 가장 큰 한계는 가상인간이 한 문장을 말할 때 같은 표정을 계속 유지한다는 것이다. 1개의 문장마다 1개의 고정된 표정을 유지하면서 말하는 것이 아니라, 실시간으로 Blendshape weight가 적절하게 바뀌면서 생동감 있는 표정을 생성하는 것이 사용자 경험을 더 향상 시킬 것이다. 이를 위한 데이터 셋이 충분하지 않아 이번 연구에서는 진행하지 못 하였다. 하지만 이런 데이터 셋이 없는 문제를 타 연구[16]에서 3D Face Tracker를 구현하여 기존 RGBD 얼굴 캡쳐 비디오에서 Blendshape을 추출해내는 방식으로 해결한 사례가 있어 이를 참고하여 향후 과제를 진행할 예정이다. 또한, 본 연구의 시스템에서는 멀티모달 대화의 구성 요소 중 제스처는 빠져 있다. 발화에 맞는 실시간 제스처 생성은 음성 인풋을 이용해 제스처를 생성하게 하는 Gesticulator[17] 등의 연구 사례가 있다. 이들을 참고하여 제스처까지 구현해 인간의 실제 의사소통과 더 가까운 가상인간의 멀티모달 대화를 구현하는 것이 목표이다.

또한 본 연구에서 사용자는 아바타 없이 텍스트로만 대화에 참여한다. 만약 현재의 시스템으로 사용자의 음성 입력을 받게 된다면, 해당 음성이 TTS로 합성되는 것이 아니기 때문에 각 음소의 타이밍을 알 수 없다. 따라서 사용자 아바타를 구현하더라도 그 아바타는 사용자 음성에 따라 적절하게 립싱크를 할 수가 없다. 타 연구에서 HMD와 STT(speech-to-text)를 이용하여 사용자의 음성 입력에 맞는 아바타의 립싱크를 구현한 사례가 있다[18]. 이를 본 연구의 시스템에 응용한다면 음성 입력을 이용하여 자연스럽게 대화하는 사용자의 아바타 또한 구현할 수 있을 것이다.

감사의 글

이 연구는 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임 (P0012746, 2022년 산업혁신인재성장지원사업)

References

[1] Wahlster, W. Dialogue systems go multimodal: The smartkom experience. In SmartKom: foundations of multimodal dialogue systems (pp. 3-27). Springer, Berlin, Heidelberg.(2006).

[2] Lee, Lik-Hang, et al. "All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda." arXiv preprint arXiv:2110.05352 (2021).

[3] Utrecht University Department of Information and Computing Sciences Virtual Worlds division IVA 2016 Tutorial September 20 (2016)

[4] Zupan, Jure. "Introduction to artificial neural network (ANN) methods: what they are and how to use them." Acta Chimica Slovenica 41 327-327.(1994).

[5] Lewis, John P., et al. "Practice and theory of blendshape facial models." Eurographics (State of the Art Reports) 1.8 2. (2014).

[6] McDonnell, Rachel, et al. "Model for predicting perception of facial action unit activation using virtual humans." Computers & Graphics 100 81-92. (2021).

[7] Ekman, Paul, and Wallace V. Friesen. "Facial action coding system." Environmental Psychology & Nonverbal Behavior (1978).

[8] Cohn, Jeffrey F., Zara Ambadar, and Paul Ekman. "Observer-based measurement of facial expression with the Facial Action Coding System." The handbook of emotion elicitation and assessment 1.3 203-221. (2007).

[9] Friesen, W. "EMFACS-7: Emotional Facial Action Coding System. Unpublished manual/W. Frisen, P. Ekman." (1983).

[10] Radford, Alec, et al. "Improving language understanding by generative pre-training." (2018).

[11] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).

[12] Li, Yanran, et al. "Dailydialog: A manually labelled multi-turn dialogue dataset." arXiv preprint arXiv:1710.03957 (2017).

[13] Rashkin, Hannah, et al. "Towards empathetic open-domain conversation models: A new benchmark and dataset." arXiv preprint arXiv:1811.00207 (2018).

[14] Zhang, Saizheng, et al. "Personalizing dialogue agents: I have a dog, do you have pets too?." arXiv preprint arXiv:1801.07243 (2018).

[15] Smith, Eric Michael, et al. "Can you put it all together: Evaluating conversational agents' ability to blend skills." arXiv preprint arXiv:2004.08449 (2020).

[16] Pham, H. X., Wang, Y., & Pavlovic, V. "End-to-end learning for 3d facial animation from speech." Proceedings of the 20th ACM International Conference on Multimodal Interaction. (pp. 361-365). (2018).

[17] Kucherenko, Taras, et al. “Gesticulator: A framework for semantically-aware speech-driven gesture generation.” In Proceedings of the 2020 International Conference on Multimodal Interaction (pp.242-250).(2020).

[18] 이재현, 박경주.대화형 가상 현실에서 아바타의 립싱크.컴퓨터그래픽스학회논문지,26(4),9-15.(2020).

〈저자소개〉

김기락
- 2021년 연세대학교 학사
- 2022년~현재 서강대학교 석사과정
- 관심분야 : Cross-Modal Generation, Metaverse
- https://orcid.org/0000-0002-2960-4583

연희연
- 2022년 서강대학교 학사
- 2022년~현재 서강대학교 석사과정
- 관심분야 : 자연어처리, 대화시스템
- https://orcid.org/0000-0003-4310-9818

은태영
- 2017년 한국외국어대학교 학사
- 2022년~현재 서강대학교 석사과정
- 관심분야 : 블록체인, 보안, 네트워크 인프라
- https://orcid.org/0000-0003-3282-8235

정문열
- 1980년 서울대학교 계산통계학과 학사
- 1982년 KAIST 전산학과 석사
- 1992년 펜실베니아 전산학과 박사
- 2018년~현재 서강대학교 아트&테크놀로지학과/메타버스전문대학원 교수
- 관심분야 : Machine Learning, VR/AR, Interactive Media Arts
- https://orcid.org/0000-0003-2462-7820