대화 맥락 기반 감정 추론을 활용한 VRoid NPC의 표정 및 제스처 표현 시스템

저자 신구대학교 - 김평오 , 윤소연

I. 서론

게임 및 가상 환경에서 NPC는 표정과 제스처 등 비언어적 요소로 감정을 전달하지만, 대화 내용과 비언어적 표현이 일치하지 않으면 사용자가 의도된 감정을 파악하기 어렵다.[1-3] 본 연구에서는 대화 맥락을 기반으로 NPC의 감정을 추론하고, 이를 가상 인간 NPC의 표정과 제스처로 시각화하는 시스템을 구현한다. 또한 "대화 감정을 반영한 NPC의 표정 및 제스처가 사용자의 감정 인식과 표현 수용에 효과적인가?"라는 질문을 감정 인식(의도된 감정을 올바르게 파악하는가)과 표현 적절성(답변·표정·제스처가 대화 맥락에 적절한가)의 두 관점에서 사용자 설문으로 평가한다.

II. 관련 연구 및 기술

NVIDIA ACE 및 Audio2Face

NVIDIA ACE는 대화형 가상 인간 구현을 위한 AI 플랫폼으로, 음성, 지능 및 애니메이션 기술을 제공한다. 이 중 Audio2Face-3D는 음성 입력을 기반으로 3D 캐릭터의 얼굴 애니메이션을 생성하며, Audio2Emotion은 음성에서 감정을 추론하여 표정에 반영한다. 해당 기술은 NIM 마이크로서비스 형태로 제공된다.

VRM Expression

VRM은 3D 아바타를 위한 모델 형식으로, Expression 기능을 통해 얼굴 표정을 표현할 수 있다. VRM 1.0에서는 표정의 강도를 0~1 범위의 가중치로 제어하도록 정의하고 있다.

본 연구와의 차별점

기존 기술은 음성 기반으로 감정을 추론해 얼굴 애니메이션을 생성하고, 선행 연구는 텍스트의 감정을 5가지로 분류해 정해진 애니메이션을 재생한다. [1,2] 본 연구는 대화 맥락에서 감정과 제스처를 독립적으로 추론하고 강도와 지속 시간을 반영하여 VRM 기반 NPC에 적용한다. 관련 기술은 참고용이며 본 시스템에 직접 사용하지 않았다.

III. 시스템 설계 및 구현

시스템 구성

본 시스템은 Unity 클라이언트, Python Flask 서버, Groq API로 구성된다. Unity에서 입력한 대사는 서버로 전달되고, 서버는 Groq API(gpt-oss-20b)로 대화 응답과 감정 분석 결과를 생성해 JSON으로 반환한다. Unity는 이 결과로 VRoid NPC의 표정과 제스처를 제어한다.

감정 추론 및 표현

감정 · 제스처

감정 6종(기쁨, 슬픔, 분노, 공포, 놀람, 중립)과 제스처 18종(인사, 동의, 생각, 부정, 울기, 웃기 등)을 사용한다. 고정 매핑 없이 AI가 대화 맥락에 따라 각각 독립적으로 추론하므로, 같은 감정이라도 상황에 따라 다른 제스처가 선택될 수 있다.

데이터 전달

분석 결과는 대화 응답(text), 감정(emotion), 강도(intensity, 0~1), 제스처(gesture), 지속 시간(duration)을 담은 JSON으로 Unity에 전달된다. 지속 시간은 대사의 음절 수, 문장부호, 감정 강도를 반영하여 서버에서 계산한다.

표현 제어

표정은 VRM Expression 가중치를 감정 강도만큼 조절하고, 제스처는 Animator로 재생한다. 새로운 대화가 들어오면 기존 표현에서 부드럽게 전환되며, 지속 시간이 지나면 중립 표정과 Idle로 돌아간다.

동작 모드

사용자 입력 기반은 사용자 대사의 감정을, AI 응답 기반은 AI가 생성한 답변의 감정을 분석하여 NPC의 표정과 제스처로 출력한다. 두 모드 모두 동일한 분석 기준을 사용하여 결과를 비교할 수 있다.

IV. 실험 및 평가

시스템 구조



[그림 1] 전체 시스템 구조도

Unity에서 입력한 대사가 Flask 서버를 거쳐 Groq API로 분석되고, 그 결과(JSON)에 따라 NPC의 표정과 제스처가 표현된다.

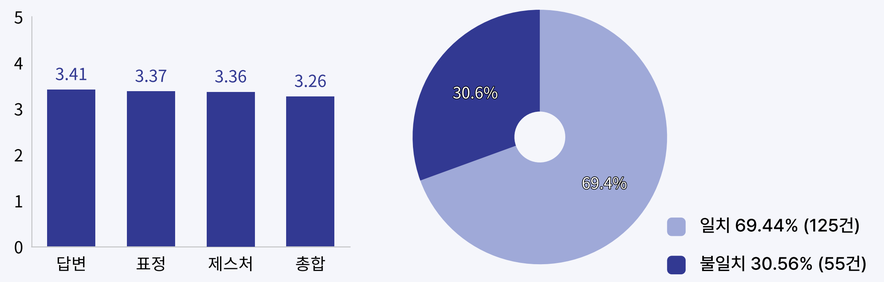
NPC와 대화 모드 실행 예시



[그림 2] NPC대화 모드 예시

"진짜 오랜만이다!! 너무 반가워!!" 입력에 대한 AI 응답 결과 (감정 Happy · 강도 0.65 · 제스처 Greeting · 지속 4.51초 · 지연 1.32초)

설문 평가 결과



표현 적절성 평균 점수 감정 인식 일치율

감정 인식 일치율 = 정답 감정 선택 응답 수 ÷ 전체 감정 선택 응답 수 × 100

결과 해석

감정 인식 일치율은 69.44%로 영상별 차이가 있었으며, 표현 적절성은 평균 3.26~3.41점으로 답변이 가장 높고 종합 평가가 가장 낮았다.

V. 결론

본 연구는 대화 맥락에서 감정 6종과 제스처 18종을 추론하여 VRoid NPC의 표정과 동작으로 표현하는 시스템을 구현하였다. 참여자를 대상으로 한 중간 설문 결과, 감정 인식 일치율은 69.44%, 표현 적절성은 모든 항목에서 평균 3.26점 이상(5점 만점)으로 나타나, 대화 맥락 기반의 감정 정보를 NPC의 비언어적 표현으로 연결할 수 있음을 확인하였다.

참고 문헌

[1] NVIDIA. (2025). Audio2Face-3D SDK (Audio2X SDK) [Computer software]. GitHub. Retrieved November 6, 2026, from https://github.com/NVIDIA/Audio2Face-3D-SDK
[2] 소훈, 정현준. (2024-11-21). LLM과 언리얼 엔진 기반의 감정 표현 NPC 시스템 구현. Proceedings of KIIT Conference, 제주.
[3] 강선경, 이민혜. (2024). ChatGPT API를 이용한 지능형 NPC 연구. 한국정보통신학회논문지, 28(4), 405-411. 10.6109/jkiice.2024.28.4.405