디지털휴먼엔터테인먼트특론 과제 1

120220210 고재현

2023년 3월 8일

1 개요

1.1 목적

- 최근 대두되는 디지털 휴먼 관련 기반 기술들을 조사한다.
- 해당 기반 기술을 구현하기 위해 핵심적인 요소들에 대해 고려해 본다.

1.2 요소 기술 및 최근 합성 기술의 동향

한유아와 같은 디지털 휴먼을 만들기 위해서는 얼굴 합성 기술, 모션 합성 기술, 음성 합성 기술이 필요하다. 사실적인 표현을 위한 캐릭터 모델링 및 렌더링 기술도 중요하지만, 인공지능 기반의 기술들만 서술하도록 한다. 최근 대두되는 생성 모델은 크게 두 가지의 흐름으로 분류할 수 있다. normalizing flow와 LM과 Transformer를 결합한 모델이 그것이다. normalizing flow는 그 이름에서 알 수 있듯이, 데이터의 분포를 정규화하는 방식으로 데이터를 생성한다. 학습 속도는 느리지만, inference 속도는 적은 수의 point를 생성하는 데에는 충분하다[1]. LM 기반의 생성 모델은 고품질의 Tokenizer 및 Transformer 기반 예측 모델의 발전으로 최근 Vall-E[2]와 같이 빠른 속도의 학습 및 생성이 가능하다.

1.3 요소 기술별 동향

1.3.1 얼굴 및 모션 합성 기술

얼굴 또는 모션 합성 기술은 주로 키포인트(랜드마크 등)를 생성한 뒤, 해당 키포인트를 기반으로 모델링을 덮어씌우는 방식으로 수행[3]되거나, 원본 영상을 변형하여 각 영상 프레임을 생성하는 방식[4]으로 수행된다. 스마일게이트 AI Media Team의 블로그 자료를 보면, 이와 관련하여 원본 영상을 변형하여 각 영상 프레임을 생성하는 방식의 실시간 처리를 시도한 바가 있다. 합성 속도 및 품질 간의 Trade-off가 큰 분야이나, NVIDIA의 모델들이 높은 품질 및 빠른 속도의 아바타를 생성하고 있다.

1.3.2 음성 합성 기술

음성 합성 기술은 VITS[5]와 같이 Transformer 기반의 모델들이 SOTA를 차지하고 있다. 특히 Vall-EX[2] 와 같은 모델은 학습 데이터 정제 없이, 적은 입력 프롬프트로도 좋은 결과를 생성한다.

2 디지털 휴먼 모델의 발전 방향

현재 디지털 휴먼의 요소 기술들은 품질 면에서(MOS Score 등) 사람과 꽤 유사한 성능을 보여주고 있다. 특히 Vall-EX[2]는 TTS 분야에서 빠른 속도와 높은 품질 및 다양한 스타일 변환을 동시에 달성하여서, 음성 합성 분야는 거의 해결된 문제라고도 볼 수 있다. 그러나 모션 합성 분야에서의 디테일한 표현이나, 학습에 매우 많은 양의 데이터가 필요한 등의 문제는 아직 남아있다. 게임 등에 디지털 휴먼을 활용하기 위해서는 다양한 버전을 생성해야 하는 경우가 생길텐데, 성격, 목소리, 외형 등을 controllable하게 조정하는 모델이 필요하다.

2.1 추후 개발 시 고려사항

- 각 요소 기술별로 어느 데이터/모델을 사용하고 있는가?
- 각 요소 기술의 현재의 controllablity는 어느 정도인가?
- E2E Integration 되어 있는 부분은 무엇이 있는가?
- Multimodal(Music, Text, Motion)을 잘 고려하고 있는가?

참고 문헌

[1] H.-K. Song, S. H. Woo, J. Lee, S. Yang, H. Cho, Y. Lee, D. Choi, and K.-w. Kim, "Talking face generation with multilingual tts," 2022.

[2] Z. Zhang, L. Zhou, C. Wang, S. Chen, Y. Wu, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," 2023.

[3] A. Richard, M. Zollhöfer, Y. Wen, F. D. la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," *CoRR,* vol. abs/2104.08223, 2021.

[4] Y. Zhou, D. Li, X. Han, E. Kalogerakis, E. Shechtman, and J. Echevarria, "Makeittalk: Speaker-aware talking head animation," *CoRR,* vol. abs/2004.12992, 2020.

[5] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *CoRR,* vol. abs/2106.06103, 2021.