입력 텍스트로부터 음성 스타일을 반영할 수 있는 가상 인간 연구

권혜준, 김세은, 노희애, 서연수, 최승호*
이화여자대학교

*jcn99250@naver.com

Study of virtual humans that can reflect speech style from input text

Kwon Hae Jun, Kim See Eun, No Hee Ae, Seo Yeon Soo, Choi Seoung-Ho *
Ewha W. Univ.

요 약

기존 가상 인간 연구에서는 가상 인간 생성에 초점이 맞추어서 연구가 되어 오고 있다. 다만, 가상 인간 경우 원하는 목소리 기반으로 음성이 생성이 되지 않는 문제가 있다. 이를 경감하고자 텍스트 입력을 기반으로 음성 스타일이 반영될 수 있는 가상인간 연구 방법을 제안한다. 먼저, 딥페이크 모델을 생성하여 후 처리 한 뒤 예측된 립싱크 모델에 병합한다. 이후 입력으로 text 를 전달하면 보이스 스타일링이 포함한 TTS 모델 그리고 립싱크 모델을 거쳐 출력으로 가상인간이 생성이 된다. 제안 방법을 검증한 결과 가상인간의 목소리에 스타일이 반영됨을 확인했다.

Ⅰ. 서 론

가상 인간 모델을 생성하는 딥 페이크 기술에 대한 연구는 여러 분야에서 진행되고 있다. 가상 인간 생성, 혹은 발화에 맞춰 입의 형태를 변형 하는 모델은 현재까지도 연구가 활발하게 이루어져 있고 있다. 립싱크 모델에서는 기존 기술을 개선하는 방향으로 연구가 되어져 오고 있다. 예를 들어 Wav2Lip 모델의 경우 2020 년 제안된 기술인 LipGA 에서 발전하여 음성 길이 어휘 개수에 제한이 없으며 더 정확한 립싱크를 생성한다[1]. 딥페이크를 활용하여 가상 인간을 생성하는 기술로 가상 튜터를 만들어 프로그램에 대한 연구도 이루어 지고 있다[2]. 다만 다수의 기존 연구들은 가상 인간 생성과 립싱크 기술에 집중한 것이 다수이며 목소리에 대한 보이스 스타일일이 이뤄지지 않는 경우가 많다는 문제점이 있다.

이를 개선하고자 특정 인물의 가상인간을 립싱크와 딥페이크 모델을 이용해서 형성하고, 생성된 가상 인간이 입력 테스트에 대해 입 모양을 맞춰 발화하며 그 인물의 목소리를 모방하는 보이스 스타일링이 반영되는 방법을 제안한다.

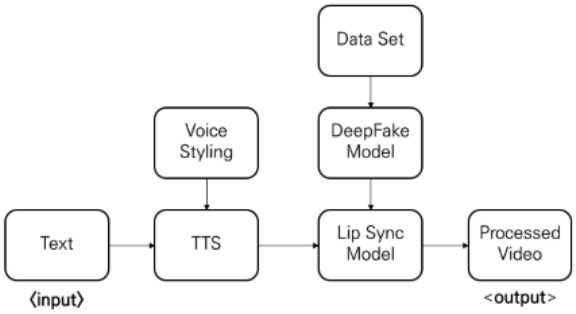

Figure 1 Our Proposal

Ⅱ. 본론

본 논문에서 제안하는 방법을 그림 1 과 같다. 먼저 TTS 에서 텍스트와 음성의 스타일을 입력 받아서 새로운 음성을 생성하게 된다. 그리고 TTS 에서 생성된 음성이 LipSync 모델에 입력으로 사용된다. LipSync 모델에서는 딥페이크에서 생성된 이미지 또한 사용된다. 그리고 LiySync 에서 생성된 가상 인간 이미지가 생성되어 나온다.

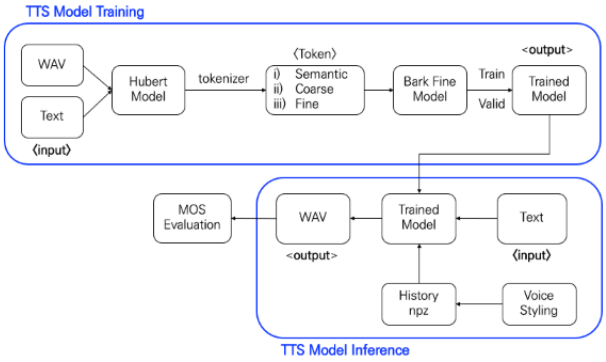

Figure 2 TTS Model Diagram

제안하는 TTS 모델은 그림 2 와 같다. 순서는 다음과 같다. 먼저 활용할 WAV 파일과 이에 대응되는 text 를 training data, validation data 로 입력한다. 이후 Hubert model 의 tokenizer 함수를 이용하여 WAV 파일을 3 종의 토큰으로 반환한다. 사용된 token 은 semantic, coarse, fine 으로 Semantic token 은 텍스트의 단어, 구문의 의미를, Coarse token 은 텍스트의 문장, 구의 리듬, 억양을, Fine token 은 텍스트의 개별 음소와 발음을 표현한다. 다음으로 Bark TTS 모델의 fine_2

모델을 이용하여 token 화 된 training data set 을 학습시킨 뒤 학습이 완료된 모델을 저장한다. 모델 학습이 완료된 이후에는 추론을 시작한다. 이때 보이스 스타일링을 위해 필요한 음성 파일을 적용하는 과정을 거친다. 활용 data 가 한국어 기반임을 고려하여 semantic history prompt npz 를 활용하는데, Bark TTS 모델에서 제공하는 함수 파일을 수정하여 준비된 인물 목소리 데이터(유재석)로 적용할 수 있게끔 한다. 또한, fine history prompt npz 도 활용할 인물에 대한 음성 npz 를 준비하여 변경하는 과정을 거친다. 이 과정에서도 history prompt npz 를 통해 text 의 음성을 활용하고자 하는 인물의 음성으로 변경하여 WAV 파일을 생성한다. 이를 통해 TTS 모델에서의 음성을 입력된 WAV 파일과 text 파일에 대응하는 인물의 목소리 생성이 가능한 보이스 스타일링 모델을 완성한다. 생성된 TTS 모델에 대해서는 MOS 를 이용하여 정량적으로 성능을 평가한다. MOS 지표의 성적은 매우 나쁨(1)~매우 좋음(5)의 범위 내에서 평가를 진행하며, 평가 인원은 총 3 인이다.

제안하는 딥 페이크 모델의 구성 순서는 다음과 같다. 먼저, 합성할 인물의 정면 얼굴 영상 data 를 이용하여 얼굴 자료 추출한다. 이후 모델 학습을 위해 SAEHD 모델을 통해서 준비한 자료 영상(target video)에서 인물 얼굴의 key point 를 추출하고 이를 학습시킨다. 최종적으로 얼굴 key point 가 학습된 파일을 바탕으로 딥 페이크 모델이 생성된다. 이후 영상 후처리 과정에서 얼굴 위치를 조정하고 blur 기능을 사용하여 합성된 영상의 질을 개선한다.

먼저 적용할 WAV 파일을 입력 받아 오디오 데이터를 20 개의 특징 벡터를 가진 Mel Frequency Cepstral Coefficients(MFCC) 형태로 변환한다. 이후 LSTM layer 와 FC layer 를 통해 20 쌍의 lip landmark 에 대한 (x, y) 좌표를 예측한다. 이때 활용할 인물 사진 data 에 대한 예측 값을 역 표준화 하여 frame 별로 (x, y) 좌표의 최솟값을 구한다. 다음으로 사용할 인물 영상(약 5 분)을 20 초씩 분할하여 frame 별로 앞서 구한 (x, y) 좌표의 최솟값에 맞춰 적용한다. 이후 data set 으로 준비한 WAV 파일을 덧씌워 영상 파일로 출력한다. 선택된 인물 영상을 덧씌울 원본 영상 데이터(5 분 분량)을 20 초간 분할하여 frame 별로 입에 대한 landmark 를 추출한다. 이후 이 landmark 의 행렬 (x, y) 좌표를 표준 화하며 데이터를 frame 화 한다.

Table 1 Mos Evaluation of TTS Model

Scale	Total	MOS Score
뜸 들이기, 끝 올려 말하는 억양	14	4.667
목소리 음역대	6	2.000
속도	11	3.667
발음	13	4.333
매끄러운 연결	10	3.333
높은 목소리 모방	14	4.667
낮은 목소리 모방	13	4.333

보이스 스타일링이 적용된 딥 페이크 모델의 데모 영상은 (Video Link)에서 확인할 수 있다. 그림 3 는 본 논문에서 제안한 방법의 데모의 일부 장면이다. 전반적인 모델의 경우 보이스 스타일링을 거친 딥 페이크 영상 형성이라는 목표를 수행하는 데에 있어서 문제는 없었으나, TTS 모델에서 음역대 모방의 정확성이 떨어지는 점, text 길이가 길어지면 부자연스러운 끊김이 발생하는 점, 또 모방의 유사성이 다소 떨어지는 점에 한계가 있었다. Lip Sync 모델의 경우 딥 페이크 처리 후 그 움직임과 텍스트에 따라 가동하는 것은 문제가 없었으나, 채색된 lip 모델을 사용하는 방식이 자연스럽지 못하다는 한계가 있었다.

Figure 3 Demo Results

Ⅲ. 결론

기존의 가상 인간 모델에 대한 연구들은 가상 인간 생성에 집중한 것이 다수이며 목소리에 대한 보이스 스타일링이 이뤄지지 않은 경우가 많았다. 이에 우리는 보이스 스타일링 기능이 추가된 TTS, 딥 페이크, Lip Sync 모델을 통해 가상 인간을 형성하는 모델을 제안한다. 실험 결과에 따르면 TTS 모델의 목소리 높낮이와 억양, 발음 부분은 높은 정확도를 보였지만 음역 대, 속도, 발화의 매끄러운 연결에 있어서는 일부 성능이 떨어진 것으로 나타났다. 이러한 결과는 제한된 영상 데이터셋과 전 처리 과정에서 외부 잡음 처리의 어려움, 그리고 WAV 파일을 작은 numpy 배열로 변환하는 과정에서 발생하는 정확도 저하 때문인 것으로 판단된다. Bark Tokenizer 를 활용한 방법 대신 멜 스펙트로그램과 같은 최신 TTS 연구에서 사용되는 CNN 기법을 적용한다면 데이터 저장의 고품질화가 가능할 것으로 기대된다.

.

참 고 문 헌

[1] Wav2lip [Internet]. Available: https://github.com/Rudrabha/Wav2Lip

[2] JaeSeong Ju, WooJin Hong, WooHyeok Moon, SaRa Yu, KyeongSoo Ko, YeongSeo Go, HaEun Jang, SiHyeon Kim, JeongHoon Shin, SeoungHo Choi, "Deep Learning-based Virtual Tutor Program to Improve English Learning Skills for Beginners in English", Journal of Digital Contents Society, Vol. 24, No. 4, pp. 1-4, Apr. 2023