AI 아나운서
: 인공지능 기술을 이용한 정보 전달 소프트웨어

김혜원*, 이영은*, 이홍창**
*울산대학교 IT 융합학과, **현대엘리베이터
e-mail: alsldjcjstk@naver.com

AI Announcer
: Information Transfer Software Using Artificial Intelligence Technology

Hye-Won Kim*, Young-Eun Lee*,
Hong-Chang Lee**
*Dept. of IT Convergence, Ulsan University
**Hyundai Elevator

요 약

본 논문은 AI 기술을 기반으로 텍스트 스크립트를 자동으로 인식하고 영상 합성 기술을 응용하여 텍스트 정보를 시각화하는 AI 아나운서 소프트웨어 연구에 대하여 기술한다. 기존의 AI 기반 영상 정보 전달 서비스인 AI 앵커는 텍스트를 인식하여 영상을 합성하는데 오랜 시간이 필요하였으며, 특정 인물 이미지로만 영상 합성이 가능했기 때문에 그 용도가 제한적이었다. 본 연구에서 제안하는 방법은 Tacotron 으로 새로운 음성을 학습 및 합성하여, LRW 데이터셋으로 학습된 모델을 사용하여 자연스러운 영상 합성 체계를 구축한다. 단순한 얼굴 이미지의 합성을 개선하고 다채로운 이미지 제작을 위한 과정을 간략화하여 다양한 비대면 영상 정보 제공 환경을 구성할 수 있을 것으로 기대된다.

1. 서론

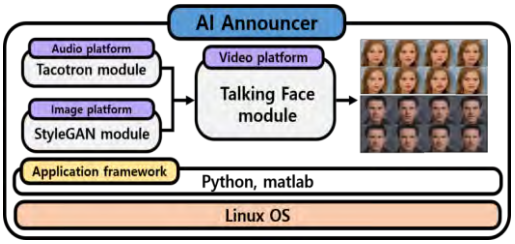
최근, AI 기술은 딥러닝(Deep Learning) 학습기법을 통하여 빠른 속도로 진화하고 다양한 분야에 응용되고 있어 곧 다가올 미래에 우리의 생활 주변에 널리 적용될 것으로 예상된다. 이러한 AI(Artificial Intelligence; 인공지능) 기술을 기반으로 실시간 영상 합성을 지원하는 AI 앵커 서비스가[1] 등장함에 따라 딥러닝을 이용한 AI 영상 기술은 전 세계적으로 많은 관심을 받고 있다. AI 앵커 서비스는 사람이 직접 영상에 출연하지 않더라도 텍스트 스크립트의 내용을 자동으로 파악하고 이를 기반으로 영상 합성을 통해 실제 아나운서처럼 표정을 짓고 말할 수 있다. 그러나 중국이나 미국과 같은 선도 국가와 함께 국내의 일부 기업에도 시도한 AI 영상 합성 시범서비스가 세계적으로 성공한 유일한 사례이며, 아직 완전한 상용화는 어려운 단계일정도로 기술적 난이도가 높은 분야이다.[2] 특히 딥러닝 학습에 필요한 데이터 수집 및 전처리 과정에 오랜 시간이 소모되고, 특정한 인물에 한해서만 AI 앵커를 만들 수 있어 다양한 인물에 대한 AI 앵커를 만들어 내기에는 많은 어려움이 따르고 있다.[3]

본 논문에서는 AI 영상 합성 기술의 핵심을 일부 간략화 하여 응용하도록 하며, 텍스트 스크립트를 입력으로 받아 영상 출력을 통한 시각화 정보 공유를 지원하는 소프트웨어를 제안한다. AI 아나운서는 Tacotron[4]기술을 활용해 입력한 텍스트 스크립트만으로도 특정인의 목소리를 낼 수 있게끔 새로운 데이터로 음성을 학습하여 새로운 TTS(Text to Speech) 음성을 생성한다. 또한, StyleGAN[5]기술로 자연스럽고 사실적인 아나운서의 얼굴을 합성하며 이 두 기술로 만들어낸 음성, 이미지 결과물을 입력으로 받아 LRW 데이터셋[6]으로 학습된 Talking Face[7]모델을 사용하여 AI 아나운서 영상을 최종적으로 생성한다.

2. 본론

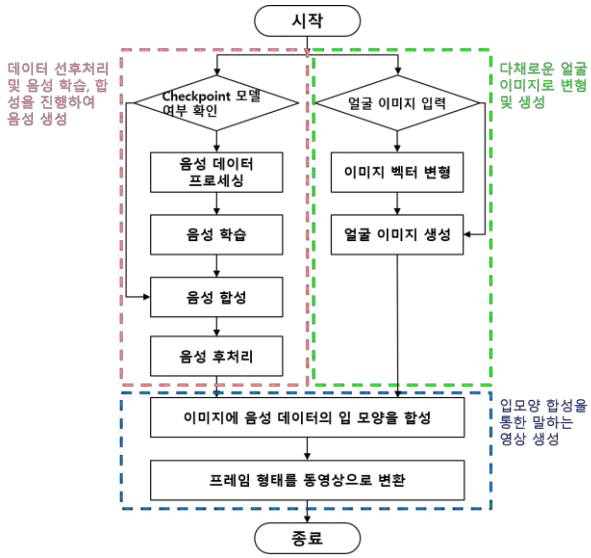
제안하는 AI 아나운서는 인공지능 기반으로 텍스트 스크립트를 자동으로 분석하고 영상 출력을 통한 시각화 정보 공유를 지원하는 소프트웨어이다. 인공지

능 음성 생성은 음성 데이터의 전·후처리 하는 모듈을 개발한 뒤 Tacotron 기술을 활용하여 학습 및 합성을 진행하였다. 텍스트와 오디오 쌍에서 직접 음성을 합성하는 방법을 학습하는 신경 텍스트 음성 변환 모델을 통해 생성된 음성파일에 원하는 텍스트 스크립트를 입력하여 학습된 목소리가 텍스트 스크립트를 읽어주는 음성 파일을 생성한다. 인공지능 가상 얼굴 생성은 StyleGAN 기술을 통해 생성된 이미지 데이터와 기존 학습된 데이터를 통해 잠재 벡터를 변형시켜 가상의 얼굴 이미지를 생성한다. 마지막으로 말하는 얼굴 생성은 Talking Face[6] 모듈을 사용해 구현하였고, 생성된 얼굴 이미지와 음성데이터를 통해 이미지에 음성 데이터의 입 모양을 합성하여 프레임을 출력 후 직접 개발한 영상 변환 모듈을 구현하여 최종 AI 아나운서 영상을 생성하였다. 그림 1 은 제안된 기술의 전체 구성도를 보여준다.



(그림 1) AI 아나운서의 시스템 구성도

2.1 시스템 흐름도



(그림 2) AI 아나운서의 시스템 흐름도

본 연구에서 제안하는 AI 아나운서 소프트웨어는 그림 2 와 같이 크게 데이터 전후처리 및 음성 학습, 합성을 진행하여 음성을 생성하는 부분과 다채로운 얼굴 이미지로 변형 및 생성하는 이미지 생성 부분, 마지막으로 앞의 두 결과물을 입력으로 받아 입 모양 합성을 통해 말하는 영상을 생성하는 비디오 부분으로 구성된다.

음성합성에 필요한 학습된 Checkpoint 모델 여부를 확인하고, 모델이 존재하지 않으면 음성학습을 진행하기 위한 데이터들을 수집한 뒤 프로세싱을 한다. 학습 파라미터들을 재설정하여 학습을 진행하고, 학습이 완료되면 생성된 Checkpoint 를 이용해 입력한 스크립트를 읽어주는 음성 합성을 한다. 그리고 영상 생성의 입력으로 들어가기 위해 음성 파일의 후처리를 진행한다. 얼굴 이미지는 입력한 이미지가 있다면 이미지의 얼굴을 인식해 나이, 성별 등 특징 벡터들을 변형하여 다양한 얼굴 이미지를 생성한다. 음성과 이미지 결과물을 입력으로 받으면 이미지의 얼굴 부위를 파악하여 음성 데이터를 통해 음성 인식 결과를 일치하며 이미지에 음성 데이터의 입 모양을 합성해 프레임 형태로 출력한다. 프레임 형태를 동영상으로 변환하고 음성을 붙여주면 말하는 AI 아나운서 영상이 생성된다. 표 1 은 제안하는 AI 아나운서의 기능 요약이다.

기능	설명
음성 학습 데이터 전처리	음성합성에 필요한 음성데이터를 만든다. API를 이용하여 음성의 텍스트를 json파일로 만든 뒤 csv 파일로 변환한다.
음성 학습	wav파일과 csv파일을 학습 가능한 numpy 형태로 변환하고, 설정한 학습 parameter에 맞춰 학습한다.
음성 생성	학습결과로 생성된 checkpoint를 불러와 입력받은 script를 읽어내는 음성을 생성한다.
음성 후처리	영상 생성 모듈에서 사용하기 위해 생성된 wav파일 음성을 bin 파일로 변환하고 알맞은 크기로 나눈다.
이미지 벡터 생성	입력받은 얼굴 이미지의 특징 embedding을 계산하여 특징 벡터를 추출하여 텐서 형태로 변환한다.
가상 얼굴 생성	이미지 벡터에 랜덤수를 넣어 매번 새로운 가상 얼굴 이미지를 생성한다.
얼굴 이미지 변형	numpy 데이터를 사용하여 얼굴 이미지의 나잇대, 성별, 표정 등을 변화하여 생성한다.
두 얼굴 이미지 합성	입력받은 두 이미지의 벡터 비중을 조절해주면서 얼굴을 합성한다.
말하는 이미지 생성	얼굴 이미지에 음성 데이터를 통해 음성 인식한 결과를 입모양에 합성하여 말하는 이미지를 생성한다.
AI 아나운서 영상 생성	생성된 프레임을 AVI비디오 변환하고, 생성한 음성을 붙여 최종 AI 아나운서 영상이 출력된다.

<표 1> 제안하는 AI 아나운서의 기능 요약

3. 구현

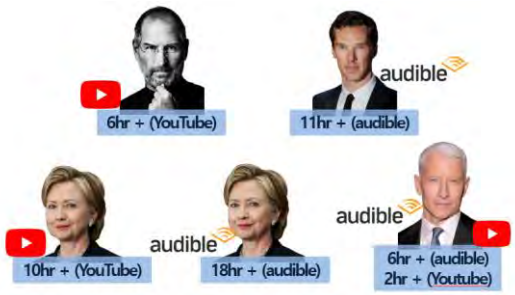
3.1 음성 생성

3.1.1 데이터 생성

Multi-Speaker Tacotron in TensorFlow[9]을 기반으로 전체음성을 학습 가능한 형태로 프로세싱하는

data_processing 모듈을 개발하고 이를 통하여 데이터를 생성하였다. 전체 음성 데이터를 pydub 라이브러리 기반의 silence.py 모듈을 통해 데이터가 묵음 구간별로 분리하고 구글 STT API를 통하여 인식 유사성이 높은 데이터만 최종 음성 데이터로 생성한다.

음성 데이터 수집은 용이한 사람들을 위주로 하여 데이터 생성을 시도하였다. 그 결과에 따라 스티브 잡스(Steve Jobs), 힐러리 클린턴(Hillary Rodham Clinton), 베네딕트 컴버배치(Benedict Cumberbatch), 앤더슨 쿠퍼(Anderson Hays Cooper) 등의 유명인을 목표로 하여 총 약 7만여 개의 음성데이터를 생성하였다. 초기 데이터 수집은 유투브(youtube), 오디블(audible) 등의 미디어를 통하여 이루어졌지만 프로세싱 후 퀄리티 저하가 심하여 최종적으로는 발화자의 목소리가 선명하게 들리는 오디오북 데이터를 프로세싱하여 힐러리 클린턴의 약 1만 5천 데이터, 앤더슨 쿠퍼의 약 5천 데이터를 생성하여 학습하였다.



(그림 3) 수집한 음성 데이터 리스트

3.1.2 음성 합성

음성합성의 기본 모듈은 Tacotron을 기반으로 하였다. 학습 시에는 빠른 시간 내에 목소리를 합성하기 위해 미리 학습된 모델에 본 연구에서 수집된 데이터를 함께 학습을 하였다. 학습 파라미터 값의 경우 sample_rate(주파수 비율)은 16000hz로 조정했으며 max_iter(최대 반복 횟수)은 최대 300 회를 넘지 않도록 조정했다. 그리고 기존 모듈에서는 learning_rate(학습 비율)은 0.002였는데 decay(음성 정보가 최고음에서 0으로 떨어지는 현상)가 1000 단계마다 진행되어 학습 단계가 지날수록 급속도로 learning_rate가 감소하였다. 따라서 본 연구에서의 모델은 learning_rate는 0.005로 설정하고 decay는 weight(가중치)를 거의 줄어들지 않게 조정하였다. 그 결과 10000 단계부터 빠르게 질이 높음 음성 데이터를 생성할 수 있었다. 또한 음성을 합성 시 텍스트 데이터는 CNN 뉴스의 스크립트를 텍스트 파일로 저장하고 자동으로 인식하여 음성을 생성하였다. 최종 결과물에서 출력되는 AI 아나운서에서 힐러리 클린턴과 앤더슨 쿠퍼는 456000 단계를 사용하였다.

3.2 이미지 생성

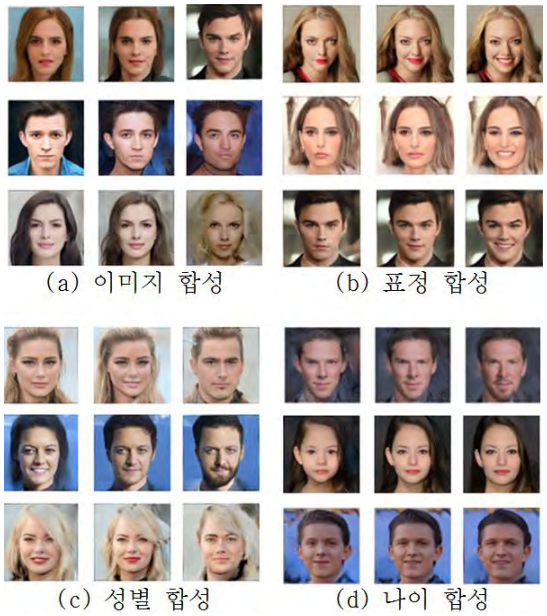
3.2.1 입력 이미지와 유사한 이미지 생성

이미지를 변환하기 위해 이미지를 numpy[8] 형태로 바꾸어서 사용한다. 먼저 입력 이미지를 bz2 모듈을 사용하여 face_landmark의 압축을 풀고 face_landmark를 사용하여 얼굴 중심으로 잘라주는 align_images.py 모듈을 사용하였다. 이 코드를 사용하면 이미지를 얼굴 중심으로 잘라주어 얼굴 변환을 용이하게 할 수 있다. 입력 이미지를 얼굴 중심으로 자른 이후에는 이미지를 변환 가능한 이미지로 생성해야 한다. 이를 위하여 오픈소스 프로젝트인 StyleGAN을 사용하여 가상의 얼굴 이미지를 먼저 생성한다. 그리고 입력 이미지와 가상 이미지를 vgg16[9] 네트워크에 삽입을 한다. vgg16 네트워크는 컨볼루션(convolution)과 최대 풀링(max pooling)을 통해 입력되는 이미지의 특징을 추출한다. 특성을 추출한 다음 각 이미지의 특성 값 오차를 구한다. 이 오차를 줄여나가기 위해 경사 하강법(Gradient descent) 알고리즘을 사용한다. 이 과정을 1000회 이상 반복하면 가상 이미지는 입력 이미지와 유사한 특징값을 가지게 된다. 이 특징값은 numpy 값을 가진다. 이미지를 제공된 age(나이), gender(성별), smile(표정)에 대한 numpy 값과 합성하여 새로운 형태의 표정을 가진 얼굴의 numpy 값을 얻을 수 있고, mix 기능에서는 두 개의 이미지를 합성하여 하나의 이미지를 만들어 낸 numpy를 생성할 수 있다.

3.2.2 StyleGAN을 사용한 이미지 합성

이미지를 생성하는 Gs_network는 StyleGAN의 이미지 생성기로 각 layer(층)마다 특정 numpy 값을 넣어 layer가 상승할수록 이미지의 형태를 띄게 되어 이미지가 생성된다. 생성된 이미지는 각 형태에 따라 변함을 보여준다. 이 이미지는 이전에 사용한 align_images.py 모듈을 사용하여 각각 얼굴 이미지로 저장할 수 있게 된다. 합성된 이미지는 numpy 값으로 이미지를 생성하여 약간의 흐릿함을 가지게 된다. 이 문제점은 opencv 라이브러리를 사용하여 이미지의 선명함을 높여주는 image_definition.py 모듈을 사용하여 흐릿한 이미지의 선명도를 높여주었다.

사용되는 인물 이미지는 바라보는 각도에 따라 결과 영상물의 얼굴형이 달라지기 때문에 눈, 코, 입 모양이 정확하게 노출되는 정면 이미지를 선정하여 실행하였다. 총 50명이 넘는 인물 이미지를 입력 이미지로 넣었고 표정, 성별, 나이 정보를 갖는 각각 100명의 인물을 생성하였고, 50명의 인물을 합성하여 100명 이상의 합성 인물을 생성하여 합성 이미지의 선명도를 높여주었다.



(그림 4) 이미지 생성 결과

3.3 영상 생성

3.3.1 AI 아나운서 프레임 생성

말하는 영상을 생성하는 모듈은 Talking-Face-Generation-DAVS를 기반으로 했다. 입력 이미지와 오디오의 포맷이 아주 민감하게 반응하는 모듈이기 때문에 앞서 처리된 음성, 이미지 데이터에 많은 전처리가 필요했다. 음성 데이터는 우리가 만들어놓은 매트랩 코드를 통해 bin 파일 xframe으로 나누어지고 mfcc_network 인코더에 입력이된다. 그리고 이미지 256x256으로 프로세싱 되고 identity_network 인코더로 입력이 된다. 이후 인코더에서 임베딩 된 값들이 Decoder에 들어가서 마지막 AI 아나운서가 프레임 형태로 출력된다.

3.3.2 AI 아나운서 영상 생성

프레임 형태로 출력이 된 AI 아나운서를 주사율(framerate)과 총 프레임 수를 계산해서 본 연구에서 개발된 img2avi 모듈을 통하여 AVI 비디오 포맷으로 출력할 수 있다. 이 출력된 영상에 ffmpeg 라이브러리를 적용하여 음성 정보가 담긴 wav 파일과 통합함에 따라 최종 AI 아나운서의 결과물이 생성된다.

(그림 5) 생성된 AI 아나운서 영상의 프레임

4. 결론

본 논문은 AI 기술을 기반으로 텍스트 스크립트를 자동으로 인식하고 영상 합성 기술의 응용하여 텍스트 정보의 시각화 정보 공유를 지원하는 AI 아나운서 소프트웨어 연구에 대하여 기술하였다. 본 연구에서 제안한 AI 아나운서는 텍스트 입력, 음성 합성, 이미지 생성, 이미지 합성, 영상 생성 단계로 이루어진다. 즉, 사용자가 텍스트와 이미지를 입력하면 각 단계를 거쳐 AI 아나운서 영상이 출력된다.

수십만건의 음성 데이터를 수집, 정제하였으며 StyleGAN 라이브러리와 딥러닝 기법을 이용하여 이미지 생성 및 합성 결과물을 출력하였다. 이를 바탕으로 Talking Face 기술을 이용하여 음성별로 합성된 이미지 프레임을 생성하였고 최종 이미지 프레임들과 음성 정보를 합성하여 음성 출력과 함께 말하는 얼굴 영상이 출력되는 AI 아나운서 시스템을 완성하였다. 차후 다양한 음성으로 TTS 서비스를 제공할 수 있도록 AI 아나운서를 확장할 예정이다.

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1] 박영진, “AI 앵커 vs 실제 앵커 과연 구분이 가능할까?,” Internet: http://www.wip-news.com/news/articleView.html?idxno=2178, Aug. 5, 2020
[2] 유하정, “국내 최초, 인공지능이 만든 ‘AI 아나운서’ 탄생!,” Internet: www.gen.or.kr/m/page/view.php?no=460, June. 20, 2020
[3] 잡플래닛, “사람 닮은 ‘AI 아나운서’, 직접 만들었죠 – 머니브레인 딥러닝팀 박성우 연구원,” Internet: www.jobplanet.co.kr/contents/news-805, Aug. 21, 2020.
[4] Yuxuan Wang, “TACOTRON: TOWARDS END-TO-END SPEECH SYNTHESIS,” Interspeech 2017, [Mar. 29, 2017].
[5] Tero Karras, “A Style-Based Generator Architecture for Generative Adversarial Networks,” CVPR 2019 final version, [Mar. 29, 2019]
[6] J. S. Chung, “Lip Reading in the Wild,” Asian Conference on Computer Vision, [Mar, 2016]
[6] Hang Zhou, “Talking Face Generation by Adversarially Disentangled Audio-Visual Representation,” AAAI Conference on Artificial Intelligence (AAAI) 2019, [Apr. 23, 2019].
[7] NumPy community, “numpy (Release 1.19.0),” Internet: https://github.com/numpy/numpy, June. 29, 2020
[8] Karen Simonyan, “VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION,” ICLR 2015, [Apr. 10, 2015]
[9] Taehoon Kim, “Multi-Speaker Tacotron in TensorFlow,”DEVIEW 2017 presentation, [Oct 15, 2017]