Journal of KIIT. Vol. 21, No. 12, pp. 67-75, Dec. 31, 2023. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2023.21.12.67

딥러닝을 이용한 텍스트 기반 개인용 콘텐츠 생성 시스템 개발

이재만*, 김선종**

Development of a Text-based Personal Content Creation System using Deep Learning

Jae-Man Lee*, Seon-Jong Kim**

본 과제(결과물)는 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 3단계 산학연협력 선도대학 육성사업(LINC 3.0)의 연구결과입니다

요 약

인공지능과 딥러닝 기술의 발전으로 인해 음성 및 얼굴합성 기술이 빠르게 완성도를 갖추어 진화하고 있으며 이러한 기술들을 활용하여 다양한 분야에서의 응용이 이루어지고 있다. 특히, 음성 및 얼굴합성 기술은 증강현실, 게임, 교육과 같은 다양한 디지털 콘텐츠에 맞게 소비되고 있으며 앞으로도 활용도가 점점 더 많아질 전망이다. 본 논문에서는 이러한 기술을 이용하여 텍스트 기반의 개인용 콘텐츠를 생성하는 시스템을 구현하였고 주어진 임의의 문장에 대해 음성과 영상에서 사용자에 맞추어 동영상으로 복구된다는 것을 알 수 있었다. 복구된 동영상의 음성은 개인적 특성이 반영되었고, 얼굴합성도 어느 정도 원활한 출력을 확인할 수 있었다. 따라서 제안한 시스템은 개인별 특성을 잘 파악해 데이터 세트를 만든다면 디지털 콘텐츠 생성에 활용될 수 있겠다.

Abstract

Due to the development of artificial intelligence and deep learning technologies, voice and facial synthesis technologies are rapidly evolving with completeness, and applications in various fields are being made using these technologies. In particular, voice and face synthesis technologies are being consumed for various digital contents such as augmented reality, games, and education, and are expected to be used more and more in the future. In this paper, a system for generating text-based personal content using this technology was implemented. It was found that for a given random sentence, the video is recovered from the voice and video to the user. The voice of the restored video reflected personal characteristics, and the face synthesis was also able to confirm the smooth output to some extent. Therefore, the proposed system can be used to create digital content if the characteristics of each individual are well identified and a data set is created.

Keywords

contents generation, text-to-speech, TTS, voice synthesis, personal speech, synthesis, deep learning

* 부산대학교 IT응용공학과 박사과정
- ORCID: https://orcid.org/0000-0002-9685-3870
** 부산대학교 IT응용공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-2070-290X

· Received: Oct. 31, 2023, Revised: Nov. 15, 2023, Accepted: Nov. 18, 2023
· Corresponding Author: Seon-Jong Kim
Dept. of IT Engineering, Pusan National University, Korea
Tel.: +82-55-350-5413, Email: ksj329@pusan.ac.kr

Ⅰ. 서 론

음성 및 얼굴합성 관련 연구는 꾸준한 발전을 이루어왔다. 특히, 딥러닝 기술의 발전과 하드웨어의 성능이 높아지고 활용처에 대한 수요가 끊이질 않고 더 자연스러운 합성을 위해서 지속적인 연구개발에 대한 필요성이 제기되어 왔다. 특히, 비교적 최근까지 질병으로 인한 비대면 사회가 지속된 이유로 오프라인 활동에 제약으로 인해 다양하게 온라인을 이용한 강의, 면접과 같은 당장에 수요를 충족시키는 한편, 출입을 자제해야 하는 상황에서의 즐길 거리를 위한 콘텐츠 역시 꾸준하게 증가하여왔다.

비대면 사회에서 텍스트를 통한 채팅은 의사 교환에 있어서 한계점이 있고 웹캠을 통한 본격적인 실시간 영상 대화는 부담감이 있다. 그래서 자신과 닮은 얼굴과 음성을 가진 영상을 만들어 강의나 기타 콘텐츠를 만들고 싶은 수요가 많아지게 되었다. 간단하게 재미 삼아 사용할 수 있는 스마트폰의 앱은 여러 가지 나와 있으나, 아직 자신만의 얼굴과 음성을 함께 모사하는 건 찾아보기 어렵다. 얼굴은 본인과 유사하게 합성해도 음성의 경우는 미리 마련된 화자의 음색으로 변경하는 정도로 현재 출시되어 있다. 또한, 이러한 온라인 콘텐츠의 발전으로 인해 다시 오프라인 활동을 하게 되었음에도 이미 하나의 문화로 자리매김하여 지속되는 현실이다.

음성합성에 대해 WaveNet[1]의 등장으로 딥러닝 기반의 음성합성 모델의 연구에 본격적이고도 혁신적인 모델로 볼 수 있고 이를 기반으로 많은 음성 모델이 발전할 수 있었다. 또한, G2P (Grapheme-to-Phoneme)[2]는 언어 처리 및 음성합성 분야에서 사용되는 개념으로 어떤 언어의 문자를 발음으로 변환하는 과정이다. Tacotron(Towards end-to-end speech synthesis)[3]은 텍스트에서 음성을 직접 생성하는 방법으로 지금의 주류가 된 End-to-End 방식의 음성합성 방법에 대해 영감을 주었다. Deep Voice 2[4]는 다양한 화자의 음성을 생성하는데 사용하는 다화자 음성 합성 모델로써, 여러 가지의 화자음성을 생성하는데 중점을 두고 있다. LwS(Listening while Speaking)[5]는 음성합성에 사용하는 접근 방식 중 하나로 입력 문장의 텍스트를 처리하여 음성합성을 생성하며 동시에 마이크 또는 오디오 입력을 통해 사용자의 실제 음성과 비교하며 피드백을 통해 발음과 억양에 생성을 조절하는 방식이다. Parallel WaveGAN[6]은 GAN기반으로 하는 빠른 오디오 파형 생성 모델로, 다중 해상도 스펙트로그램을 활용해 고품질 음성 생성에 중점을 둔 모델이며 본 논문에서 사용하는 모델이 기반을 두고 있다. Glow-TTS[7]는 텍스트에서 음성을 생성하는데 Generative Flow모델을 활용하고 모노토닉 정렬 검색을 통해 음성을 생성하며 고품질의 자연스러운 음성을 생성하는데 중점을 두고 있다. HiFi-GAN[8]은 효율적이면서 고품질의 음성합성을 위해 GAN을 기반으로 하는 모델로, 더 자연스러운 합성음성을 목표로 하고 있다. 이렇듯 음성합성 모델을 활용한 예로써 한국어 가요 음성합성[9]과 웹서비스에 관한 연구도 있다.

얼굴합성으로는 Face2Face[10]으로써 실시간 영상에서 얼굴표정을 획득하여 비디오로 재현하는 기술을 제안하고 있다. 또한, 생체 데이터로 음성과 얼굴 두 가지 데이터를 활용하여 사용자의 신원을 확인[11]으로 활용되기도 한다. 마지막으로 음성기반에 얼굴합성 모델로는 FaceMormer[12], GeneFace[13]이 있는데 음성입력에 따라 입술 및 얼굴표정을 생성해주는 모델로 가장 최근까지 공개된 모델로 연구가 진행되고 있다.

본 논문에서는 인공지능을 이용한 연구를 기반으로 텍스트 입력 음성 및 얼굴합성 모델을 이용하여 개인적인 음성과 얼굴을 합성하는 시스템을 구현한다. 동영상 콘텐츠는 사전에 수집된 데이터를 이용하여 성능을 평가하고, 이를 통해 사용자에 맞춰진 음성 및 얼굴을 생성할 수 있는 통합된 시스템의 기반이 될 수 있도록 한다. 사용자에 맞춰진 음성 및 얼굴 영상 데이터를 수집하고, 학습하여 새로운 텍스트 기반 콘텐츠를 생성하여 새로운 서비스를 제공할 수 있도록 한다.

2장에서는 제안하는 시스템과 음성 및 얼굴합성 모델, 그리고 데이터 수집과 평가지수를 설명한다. 3장에는 개인적인 데이터를 이용하여 주어진 인공지능 모델에 적용하여 학습한 후, 테스트 결과에 대한 고찰 및 성능평가 그리고 마지막 4장에서는 성능에 관한 결과를 기반으로 결론을 맺는다.

II. 개인용 콘텐츠 생성 시스템

그림 1은 제안된 시스템의 개인용 콘텐츠 생성 시스템을 나타낸 것이다. 우선, 사전에 개인의 영상 데이터를 기반으로 만들어진 데이터베이스에서 음성과 이미지로 분리하여 각각 TTS 및 얼굴을 학습한다. 학습이 완료되었다면 사용자가 텍스트를 입력하게 되면 학습된 맞춤형 화자의 목소리를 TTS가 합성하고 합성된 음성데이터를 얼굴합성 모델을 통해 음성의 신호에 맞도록 각각의 이미지로 얼굴을 시뮬레이션하여 최종적으로 만들어진 음성과 이미지들을 하나의 동영상으로 결합하여 완성한다.

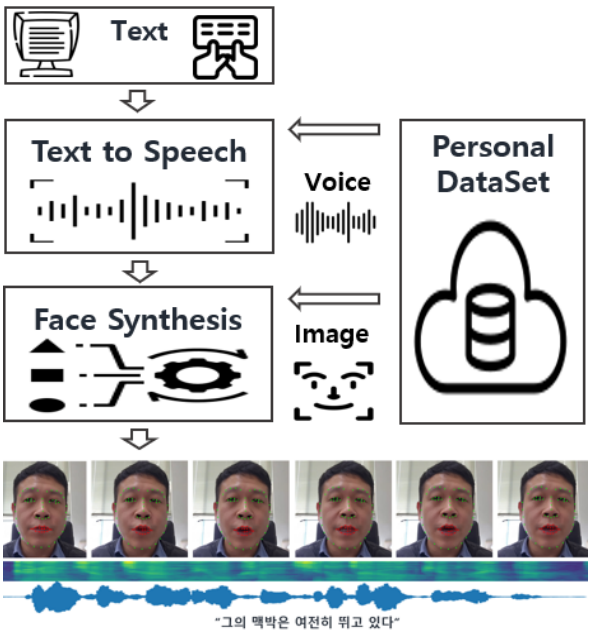

그림 1. 개인용 콘텐츠 생성 시스템
Fig. 1. Personal content generation system

2.1 TTS(Text to Speech)

본 논문에서는 앞서 서론에서 언급한 음성합성 모델들 중에서 비교적 최근에 나왔고 성능이 좋은 Glow-TTS와 HiFi-GAN을 사용하였다.

Glow-TTS는 Generative Flow for Text-to-Speech로써 기존에 FastSpeech모델이 병렬적으로 합성하여 속도를 높인 보코더를 제안했듯이 Glow-TTS는 플로우 기반 생성모델과 동적 프로그래밍의 속성을 활용해서 기존모델보다 장문의 텍스트를 합성하고 억양의 강세를 갖춘 음성에 생성이 가능하고 합성 속도 역시 개선된 모델이다. 후속적으로 사용하는 HiFi-GAN은 GANs(Generative Adversarial Networks)을 기반으로 한 음성합성 모델로 고성능 및 고품질의 음성합성을 위해 설계되었고 실제 음성과 매우 유사한 음성을 생성할 수 있다. 즉, Glow-TTS는 학습 대상인 화자의 말투에 영향을 많이 받고 HiFi-GAN은 Glow-TTS로 생성한 음성을 실제 화자의 음색에 더 가깝게 한다.

2.2 Face synthesis

최근 여러 가지 얼굴합성 모델들이 있으나, 몇 가지를 비교해 볼 때, 얼굴합성에 사용할 모델로 GeneFace를 사용한다. 장점으로는 다른 모델들에 비해 학습 시간이 비교적 짧고 합성된 얼굴의 결과물이 음성을 발음할 때 얼굴의 움직임이 학습에 사용한 영상데이터의 기반으로 약간씩 움직여 주어 자연스러움을 더해준다는 장점이 있다.

GeneFace는 오디오 기반의 입력에서 HuBERT 특징을 추출하여 머리와 몸통을 NeRF(Neural Radiance Fields)를 생성해 연속적인 얼굴 애니메이션을 생성해주는 모델로써, 음성합성 및 얼굴 모델링을 결합한 고도화된 모델이다. 해당 모델은 RAD-NeRF (Real-time neural talking portrait synthesis) 기반 렌더러로 실시간으로 추론하고 학습에 사용하는 영상데이터의 길이에 따라 데이터를 전처리와 하드웨어의 성능에 따라 차이는 있겠지만 10시간 내외로 학습할 수 있었다.

2.3 데이터 수집과 성능평가

본 논문에서는 복원하고자 하는 대상인 화자의 영상과 음성데이터를 동시에 수집한다. 이때 대본 문장을 사전에 준비하고, 준비된 대본을 한 문장씩 읽는 영상을 저장하였다.

성능평가에 사용되는 문장은 학습되지 않는 문장을 미리 선별하여 발음하는 음성과 영상을 사전에 수집하였다. 이후 같은 문장을 입력하여 합성된 음성과 영상을 가지고 평가하였다.

음성에 대한 성능평가는 F0(Foudamental frequency)와 에너지, 두 가지 값을 사용한다. 그리고 영상에 대한 성능평가는 입술에 대한 특징인 기준점(Landmarks)의 위치로 평가하였다. 원본 영상 및 합성 영상의 전체 프레임에서 두 영상의 입술 위치가 다르기에 각각의 프레임별 입술 좌표점의 무게 중심을 기준으로 정렬하였다. 또한, 원본 영상과 합성 영상의 프레임을 모두 25fps로 맞추고 문장을 읽고 끝나는 시간이 음성과 영상 모두 똑같지 않으므로 데이터 수는 둘 중 적은 쪽을 기준으로 삼는다. 정렬된 두 데이터 간에 차이를 MAE(Mean Absolute Error)을 성능평가 지수에 사용하였다. MAE는

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \quad (1)$$

이 된다.

III. 실험 결과 및 고찰

본 논문에서는 실험을 위해 2명의 화자에게서 그림 2와 같이 각각 1,000개의 문장을 읽는 음성이 포함된 동영상을 녹음하고 음성은 별도로 추출하여 음성합성에 학습하고 얼굴합성의 학습에는 영상데이터를 하나로 합쳐서 진행하였다. 이때 영상에서 발음하는 얼굴의 움직임과 크기가 많이 차이 나는 일부의 데이터는 성능에 영향을 끼치므로 포함하지 않았다.

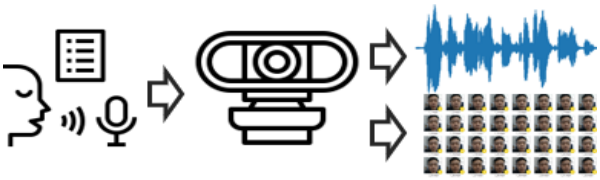

그림 2. 화자로부터 음성 및 영상데이터를 획득
Fig. 2. Obtaining video and voice data from the speaker

그림 3은 영상데이터와 문장 일부분을 나타낸 것으로 수집된 데이터에서 음성은 영상과 별도로 분리하여 학습데이터로 사용하였다.

음성과 얼굴합성에 대한 인공지능 모델의 학습이 모두 끝나고 나면 개인용 콘텐츠에 활용할 텍스트를 입력하면 만들고 싶은 화자의 음성과 얼굴이 합성되어 출력되도록 하였다. 물론 실제 사용자와 흡사한 음성합성 결과를 출력하게 되고 다시 얼굴합성 모델에 입력하여 얼굴합성을 하게 된다. 최종적으로 결과물은 동영상으로 만들어지며 그림 4과 같은 인터페이스로 제작하여 확인하였다.

그림 3. 영상데이터와 대본의 일부분
Fig. 3. Video data and some sentences

그림 4. 텍스트 기반 개인용 콘텐츠 시스템
Fig. 4. Text-based personal content system

표 1은 음성과 얼굴합성의 결과를 비교하기 위한 테스트 문장의 목록으로 10개의 문장으로 구성되어 있으며, 사전에 음성과 얼굴을 해당 문장대로 읽어 저장해 사용하였다.

표 1. 테스트 문장 목록
Table 1. List sentences for the test

No.	Sentence
1	그의 맥박은 여전히 뛰고 있다
2	한글은 15세기에 세종대왕에 의해 만들어졌어
3	그가 회복된 건 순전히 그 여자의 간호에 의한 거야
4	일기예보에 의하면 날씨가 좋을 거래요
5	IMF는 뭘 의미합니까
6	여름철에는 매미가 우는 걸 쉽게 볼 수 있어요
7	전화 받아 전화벨 울리잖아
8	오늘은 운동하기에 아주 좋은 날씨였다
9	비 오는 날 빨리 운전하면 매우 위험하다
10	그는 엉뚱한 질문으로 사람들을 곧잘 웃겼다

그림 5는 화자에 대한 음성학습에 대한 손실률을 나타낸 예로써 그래프를 살펴보면 Glow-TTS의 경우 Eval과 Train의 평균 손실이 학습의 스텝이 300k 선에서 큰 변화를 보이지 않음을 알 수 있다. HiFi-GAN의 Eval과 Train의 손실률을 보면 Eval은 수치가 0을 향해 가는 게 뚜렷하게 보이지만 Train은 경우 Average Discriminator Loss의 손실은 커지는 것을 알 수 있다. 이것은 학습이 오래 지속됨에 따라 성능이 오히려 하락함을 알 수 있다. 원인으로는 음성데이터의 수가 부족한 걸로 보이며 앞으로 화자의 음성데이터를 추가로 수집하여 성능을 향상할 여지를 가지고 있다. 현재는 두 가지 음성 모델의 학습 중 가장 좋은 지점은 약 300k 지점에서 생성된 스텝에서의 학습 파일을 사용하였다.

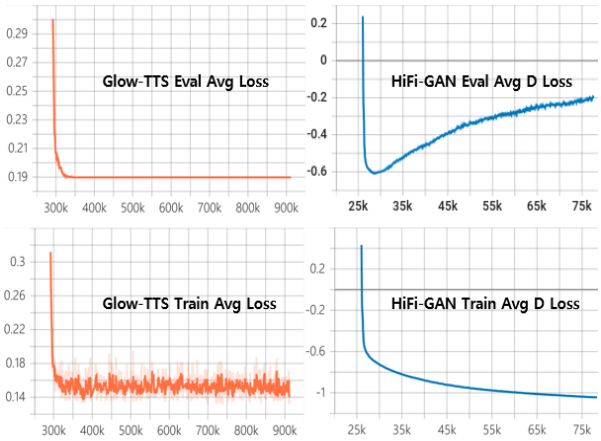

그림 5. Glow-TTS, HiFi-GAN의 Eval, Train의 결과
Fig. 5. Result of Glow-TTS, HiFi-GAN's Eval, Train

10개의 테스트 문장에 대한 합성된 음성은 화자의 특성을 살려서 어느 정도 합성되어 비슷하다는 느낌을 받을 수 있었다. 정확한 평가를 위해 그림 6, 7에 원본 음성과 합성으로 생성된 음성의 F0과 에너지를 그래프로 표현하였다. F0의 경우, 음성 신호의 음높이를 결정하는 특징이며, 에너지는 음성 신호의 강도를 나타내는 척도이다. 그래프를 보면 원본과 비슷한 구간과 그렇지 않은 구간을 확인할 수 있다.

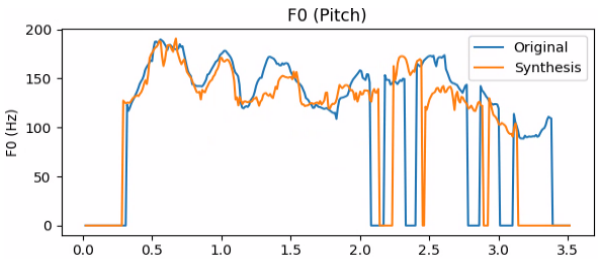

그림 6. 문장 No. 7에 대한 화자 1의 F0
Fig. 6. F0 in Speaker 1 sentence No. 7

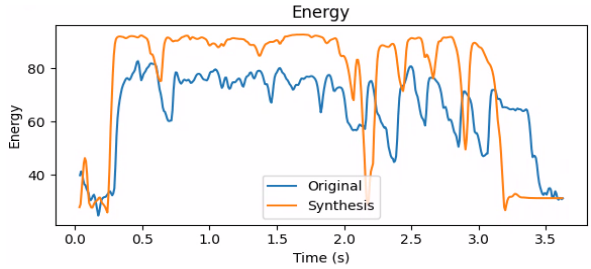

그림 7. 문장 7에 대한 화자 1의 에너지
Fig. 7. Energy of No. 7 sentence in Speaker 1

표 2는 음성합성의 평가 결과이다. 결과를 보면, F0의 차이가 에너지보다 크다는 것을 알 수 있는데, 그림 6에서 보는 것과 같이 음성 발음에 대한 차이로 인하여 오차가 발생함을 알 수 있다. 따라서 원본과 합성된 음성의 발음과 말하는 속도에 따른 문제를 해결하여야 더 정확한 평가가 될 수 있다. 그리고 에너지는 전체적인 오차는 많은 차이가 없지만, 그림 7에서 보듯이 합성된 음성에서 에너지가 많다는 것을 알 수 있다.

표 2. 화자 1, 2에 대한 F0, 에너지에 대한 성능 결과
Table 2. Performance results of F0, Energy for speaker 1 and 2

Sentence	Person 1		Person 2	
	F0(Hz)	Energy (dB)	F0(Hz)	Energy (dB)
1	48.7	7.47	73.4	11.2
2	52.2	4.43	43.6	12.8
3	35.0	2.43	63.6	12.8
4	22.2	4.08	55.1	15.5
5	42.4	7.40	53.0	12.9
6	30.5	11.9	61.4	12.9
7	46.7	12.8	44.6	12.1
8	44.2	8.78	54.9	13.5
9	23.4	5.76	52.4	12.8
10	61.9	9.57	51.7	12.6
Avg.	34.53	7.23	50.20	12.94

표 3은 F0 데이터에서 음성이 없는 부분을 똑같이 제외하고 화자 1, 2에 대해 F0의 MAE를 측정한 결과이다. 결과를 보면, 앞서 측정한 결과보다 손실률이 하락하였고 복구된 음성이 화자의 특성 차이가 축소됨을 확인할 수 있다.

표 3. 화자 1, 2에 대한 F0의 일부 구간의 성능평가 결과
Table 3. Performance results of partial F0 for speaker 1 and 2

Sentence	Person 1	Person 2
	F0(Hz)	F0(Hz)
1	27.11	22.46
2	21.19	22.16
3	19.23	34.81
4	14.16	41.14
5	31.00	42.84
6	12.81	22.37
7	19.97	36.58
8	27.36	32.06
9	16.74	46.36
10	22.54	29.77
Avg	18.96	30.08

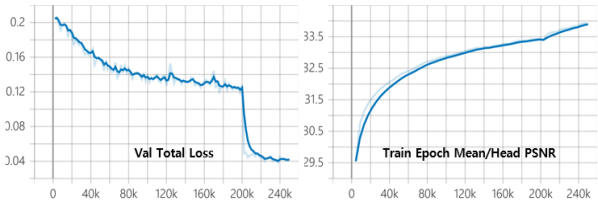

그림 8. 얼굴 영상의 검증과 학습에 대한 Loss
Fig. 8. Loss of validation and train in the face image

그림 8은 화자 1에 대한 검증 및 학습에 대한 그래프이다. 약 250k의 스텝만큼 학습하였으며 Val의 손실률이 꾸준하게 덜어지고 Train의 PSNR(Peak signal to noise ratio)이 상승하는 것을 볼 수 있는데 PSNR은 비디오의 품질을 측정하기 위한 지표 중 하나로써 원본 신호와 잡음 사이에 관계를 나타내며 비교적 학습이 양호한 것으로 보인다.

그림 9는 68개의 기준점 중에서 입술에 해당하는 20개의 좌표점을 프레임별로 변화하는 모습으로써 문장을 입으로 발화하는 과정을 나타낸다.

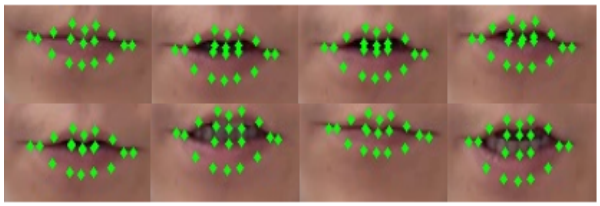

그림 9. 프레임별 입술 좌표점의 변화
Fig. 9. Changes in lip coordinate points by frame

얼굴합성의 성능을 평가하기 위해 원본 영상과 텍스트 입력으로 합성된 영상의 차이를 측정하기 위해서 우선, Dlib의 68개의 입술에 해당하는 49번부터 68번까지의 20개의 좌표점을 영상의 전체 프레임에서 추출하여 그림 10에 좌표점들을 표시하였다. 또한, 원본과 합성 영상에서 추출한 입술 좌표점이 서로 위치가 달라서 입술 좌표점의 무게 중심을 기준으로 정렬하였다.

그림 10을 살펴보면 원본과 합성의 좌표점들을 보면 문장의 발음 과정의 움직임을 알 수 있는데 실내에서 혼자 발음하는 경우 목소리를 크게 할 필요성이 적어 입을 크게 벌리지 않게 된 이유로 보이고 이것은 앞으로 학습데이터를 수집할 때 조정이 필요하다.

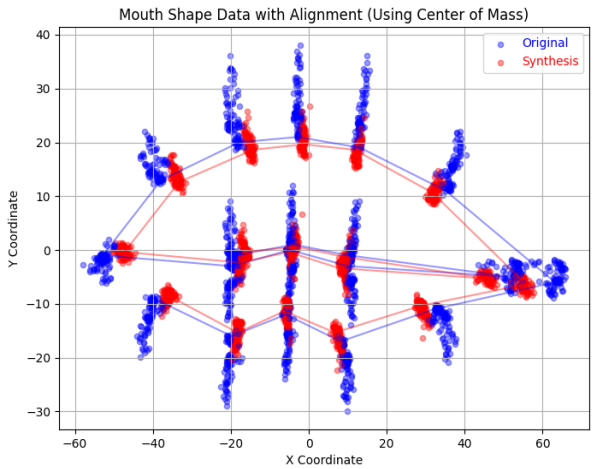

그림 10. 원본과 복구 영상에서 입술 랜드마크 위치
Fig. 10. Position of the original and synthesis lip landmarks in the video

표 4. 얼굴에 대한 MAE
Table 4. MAE of the face landmarks

Sentence	Person 1	Person 2
1	4.21	1.90
2	3.22	2.51
3	3.61	2.30
4	3.83	1.98
5	3.66	2.69
6	3.24	1.98
7	3.38	1.86
8	3.07	2.62
9	3.56	2.38
10	4.11	2.32
Avg.	3.59	2.25

표 4는 화자 1, 2에 대한 원본과 합성된 얼굴 입술에 대해 차이인 MAE를 나타낸 결과이다. 결과를 분석해 보면 음성과 마찬가지로 두 영상에서의 발음에 대한 시간이 같지 않아 움직임의 영역에서 차이가 나타남을 알 수 있었다.

전술한 바와 같이 복구된 동영상의 음성은 개인적 특성이 반영되었고, 얼굴합성도 어느 정도 원활한 출력을 확인할 수 있었다. 또한 복구된 동영상을 이용하여 음성과 영상에 대한 성능을 각각 평가한 결과, 서로 간의 발음 시간 차이로 인하여 차이가 발생한다는 것을 알았다. 이는 평가에 사용한 원본 음성과 영상은 화자의 상태에 따라 변하기 쉽기 때문이다. 또한 합성을 통해 복구된 음성과 영상의 경우는 화자의 다양한 상태의 특성을 가질 수 있다는 것을 알았다.

Ⅳ. 결 론

본 논문에서는 고품질의 음성 및 얼굴합성이 가능한 Glow-TTS and HiFi-GAN, GeneFace모델을 사용하여 텍스트 기반의 개인용 콘텐츠 생성 시스템을 제안하였다. 이는 활용하고자 하는 화자를 대상으로 음성과 영상데이터를 수집하였고 이를 인공지능에 학습하였다. 학습된 모델에 문장을 입력하여 음성 및 얼굴을 생성할 수 있는 시스템을 구현하였다. 주어진 임의의 문장에 대해 음성과 영상에서 사용자에 맞추어 동영상으로 생성된다는 것을 알 수 있었다. 합성된 음성에는 개인적 발음 특성이 잘 반영되었고, 얼굴합성도 원활하게 콘텐츠가 생성되는 것을 확인할 수 있었다. 성능평가에서는 발음 시간이 같지 않고 화자의 데이터 세트가 다양한 특성이 있어서 차이가 나타났다. 따라서 원하는 화자의 데이터 특성을 잘 고려한다면 정밀한 콘텐츠를 생성할 수 있겠다.

References

[1] A. Oord, et al., "WaveNet: A Generative Model for Raw Audio", arXiv:1609.03499, Sep. 2016. https://doi.org/10.48550/arXiv.1609.03499.

[2] S. M. Kang and M. S. Chang, "The Suggestion to Improve Performance of G2P of Korean TTS according to Collection and Analysis on Pronunciation Error Data", The Journal of Korean Institute of Information Technology, Vol. 8, No. 8, pp. 205-212, Aug. 2010.

[3] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, and B. Ramabhadran, "Tacotron: Towards End-to-End Speech Synthesis", Interspeech 2017, pp. 4006-4010, Aug. 2017. http://dx.doi.org/10.21437/Interspeech.2017-1452.

[4] S. Arik, G. Diamos, A. Gibiansky, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-Speaker Neural Text-to-Speech", arXiv:1705.08947, May 2017. https://doi.org/10.48550/arXiv.1705.08947.

[5] A. Tjandra, S. Sakti, and S. Nakamura, "Listening while Speaking: Speech Chain by Deep Learning", 2017 IEEE Automatic Speech Recognition and Understanding Workshop, Okinawa, Japan, Dec. 2017. https://doi.org/10.1109/ASRU.2017.8268950.

[6] R. Yamamoto, T. Toda, and H. Banno, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram", ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, May 2020. https://doi.org/10.1109/ICASSP40776.2020.9053795.

[7] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search", arXiv:2005.11129, May 2020. https://doi.org/10.48550/arXiv.2005.11129.

[8] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis", arXiv:2010.05646, Oct. 2020. https://doi.org/10.48550/arXiv.2010.05646.

[9] J. Park and J. Kim, "Development and Research of a Web Service for Korean Song Speech Synthesis", Journal of the Korea Information Technology Society, Vol. 20, No. 1, pp. 181-189, Jan. 2022. https://doi.org/10.14801/jkiit.2022.20.1.181.

[10] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time Face Capture and Reenactment of RGB Videos", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016. https://doi.org/10.1109/CVPR.2016.468.

[11] J. Y. Park and S. H. Ok, "A Study on a Identification Method using Multiple Data Synthesis Algorithms and Convolutional Neural Networks", The Journal of Korean Institute of Information Technology, Vol. 19, No. 11, pp. 99-106, Nov. 2021. http://doi.org/10.14801/jkiit.2021.19.11.99.

[12] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-Driven 3D Facial Animation with Transformers", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans Louisiana, pp. 18770-18780, 2022. https://doi.org/10.1109/CVPR.2022.6789477.

[13] Z. Ye, Z. Jiang, Y. Ren, J. Liu, and Z. Zhao, "GeneFace: Generalized and High-Fidelity Audio-Driven 3D Talking Face Synthesis", arXiv:2301.13430, Jan. 2023. https://doi.org/10.48550/arXiv.2301.13430.

저자소개

이 재 만 (Jae-Man Lee)

2011년 8월 : 부산대학교
바이오정보전자전공(공학사)
2014년 2월 : 부산대학교
IT응용공학과(공학석사)
2021년 3월 ~ 현재 : 부산대학교
IT응용공학과 박사과정
관심분야 : 신호 및 영상처리
머신/딥러닝

김 선 종 (Seon-Jong Kim)

1996년 8월 : 경북대학교
전자공학과(공학박사)
1995년 2월 ~ 1997년 2월 :
순천제일대학 전임강사
1997년 3월 ~ 현재 : 부산대학교
IT응용공학과 교수
관심분야 : 신호 및 영상처리,
머신/딥러닝, VR/AR, 스마트 카메라