음성 발화 기반 3차원 얼굴 메쉬 복원 모델 구현

김성진, 최문경, 허정우, 최성화, 이상훈
연세대학교
{jin.k, bryan1302, gjwjddn9, csh0772, slee}@yonsei.ac.kr

3D Facial Mesh Reconstruction Model Implementation based on a Voice Speech

Kim Seongjean, Choi Moonkyeong, Huh Jungwoo, Choi Seonghwa, Lee Sanghoon
Yonsei Univ.

요 약

음성 발화에는 발화자의 성별, 나이, 입 모양 등 발화자의 얼굴에 대한 다양한 정보가 결합하여 있다. 기존 연구들은 음성 발화에서 발화자 얼굴 이미지, 얼굴형, 움직임 등을 각각 복원하였다. 본 논문에서는 짧은 음성 발화를 기반으로 3차원 얼굴 메쉬와 그 움직임을 복원하는 시스템을 제안한다. 3차원 얼굴 모델은 FLAME (Face Learned with an Articulated Model and Expressions) 을 활용하였다. 구현한 모델을 통해 복원된 얼굴 메쉬는 가상현실 및 증강현실의 다양한 어플리케이션에서 활용될 것으로 기대한다.

Ⅰ. 서 론

사람은 음성 발화를 듣고 발화자의 얼굴을 예측할 수 있다. 이것은 발화자의 성별, 나이, 입 모양 등 물리적인 요인이 발화자의 얼굴과 함께 음성 발화에 영향을 미쳤기 때문이다. 또한 발화자의 국적, 인종, 발음 등 문화적인 요인이 발화자의 얼굴과 동시에 음성 발화와 연관성을 가지고 있기 때문이다. 이를 통해 음성 발화에는 발화자의 얼굴에 대한 다양한 정보가 결합하여 있음을 알 수 있다.

기존 연구들은 음성 발화에서 발화자의 얼굴 이미지[3,4], 움직임[1], 얼굴형[5] 등의 특성을 각각 복원하였다. 얼굴 이미지는 발화자의 성별, 나이에 대한 정보를 담고 있지만 얼굴의 움직임과 다른 방향에서 바라본 얼굴에 대한 정보를 담고 있지 않다. 반면 얼굴의 움직임과 얼굴형은 발화자의 성별, 나이에 대한 정보를 담고 있지 않다. 우리는 음성 발화를 통해 복원한 특성들을 결합하여 발화자에 대한 다양한 정보를 함께 담고 있는 3차원 메쉬를 복원하고자 한다. 본 논문에서는 3차원 얼굴 모델 FLAME (Face Learned with an Articulated Model and Expressions)을 이용하여 짧은 음성 발화를 기반으로 3차원 얼굴 메쉬를 복원하는 모델을 제안한다.

Ⅱ. 본론

3차원 얼굴 메쉬 복원 모델에서 3차원 얼굴 메쉬를 나타내기 위해 FLAME(Face Learned with an Articulated Model and Expressions)[6] 모델을 활용하였다. FLAME 모델은 머리와 얼굴을 기존 소프트웨어와 호환되며 정확하게 표현하기 위한 모델로 얼굴형, 포즈, 표정에 대한 파라미터에 따라 얼굴 메쉬를 생성한다. 얼굴형은 CAESAR 데이터셋, 포즈와 표정은 D3DFACS 데이터셋을 사용해 학습했다. 3차원 인간형 모델인 SMPL (Skinned Multi Person Linear model)과 결합해서 3차원 아바타를 쉽게 생성할 수 있다.

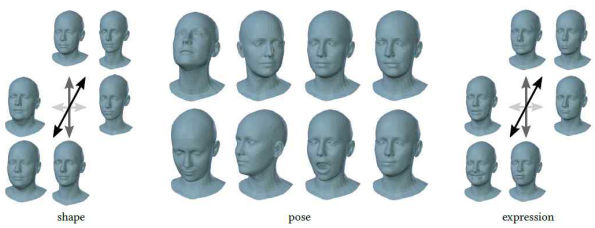

그림 1 FLAME (Face Learned with an Articulated Model and Expressions)

제안하는 모델은 발화자의 얼굴 이미지, 얼굴형, 움직임을 각각 복원하는 모델들을 결합하여 3차원 얼굴 메쉬를 복원한다. 음성 발화와 3차원 얼굴 메쉬가 있는 대규모 데이터셋이 존재하지 않기 때문에, 얼굴형 복원 모델을 학습시키기 위해 AVSpeech[2] 데이터셋을 확장시켜 음성 발화와 얼굴형 파라미터 페어가 존재하는 데이터셋을 생성하였다. 한 사람의 얼굴과 목소리가 담긴 유튜브 비디오에서 얼굴 이미지와 음성 발화를 내려받았다. 그리고 얼굴 이미지에서 FAN(Face Alignment Network)[7]를 이용해 얼굴 랜드마크를 추출하였다. FLAME 모델을 기반으로 생성한 얼굴 메쉬의 랜드마크가 이미지에서 추출한 얼굴 랜드마크와 일치하도록 얼굴 메쉬의 파라미터를 최적화하였다. 음성 발화는 비디오의 앞부분 6초의 소리를 사용하는데 만약 소리가 6초보다 짧다면 6초가 되도록 반복시켰다. 8000개의 유튜브 비디오를 이용하여 음성 발화와 얼굴형 파라미터 페어 데이터셋을 만들고 얼굴형 복원 모델을 학습시켰다.

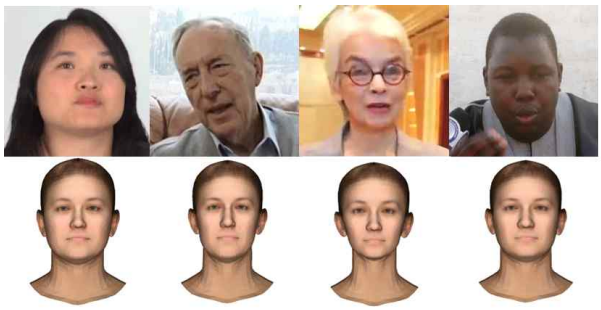

그림 2 생성한 데이터셋 예시

얼굴형 복원 모델(shape encoder)을 학습시키는 파이프라인이 그림 3에 나타나 있다. 학습 과정에서 먼저 음성 발화의 특징을 추출하기 위해 음성 발화를 스펙트로그램으로 변환하였다. 다음으로 얼굴형 복원 모델이 스펙트로그램을 기반으로 발화자의 얼굴형 파라미터를 추측하였다. 우리가 생성한 데이터셋의 얼굴형 파라미터를 정답으로 사용하여 얼굴형 복원 모델이 같은 파라미터를 출력하도록 학습시켰다. 모델의 구조는 Speech2Face[4]의 voice encoder의 구조를 사용하였다.

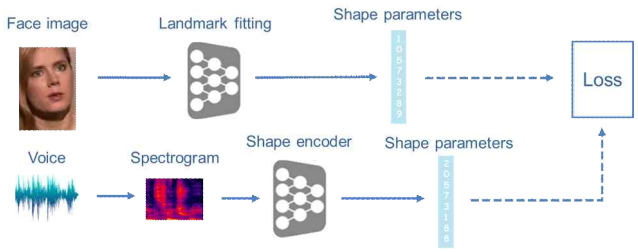

그림 3 얼굴형 복원 모델 학습 파이프라인

3차원 얼굴 메쉬 복원 모델의 파이프라인이 그림4에 나타나 있다. 얼굴형 복원 모델이 음성 발화로부터 얼굴형 파라미터를 추측하였다. 추측한 얼굴형과 일반적인 자세, 표정을 가지는 템플릿 메쉬를 FLAME 모델을 이용해 생성하였다. 템플릿 메쉬와 FaceFormer[1]를 통해 음성 발화에 맞는 움직임을 가지는 메쉬를 추측하였다. Speech2image[3]로 짧은 음성 발화로부터 저화질의 얼굴 이미지를 복원하고 GFPGAN으로 이미지를 고화질로 복원하였다. DECA를 사용해 이미지를 텍스쳐로 만들고 메쉬와 결합해 최종 결과물을 생성하였다.

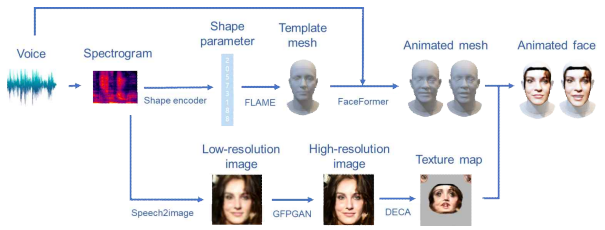

그림 4 3차원 얼굴 메쉬 복원 모델 파이프라인

복원된 얼굴 메쉬를 확인하기 위해 Voxceleb2 테스트 데이터셋을 사용하였다. Voxceleb2는 유튜브에 있는 인터뷰 비디오 15만개로 이뤄진 데이터셋이다. 발화자 얼굴 이미지와 복원된 얼굴 메쉬가 그림 5에 나타나 있다. 발화자의 실제 얼굴과 복원된 얼굴 메쉬의 성별, 나이 등의 특성이 공통적이라는 것을 확인할 수 있다. 그림 6을 통해 발화자의 실제 얼굴과 복원된 얼굴 메쉬의 움직임이 공통적이라는 것을 확인할 수 있다.

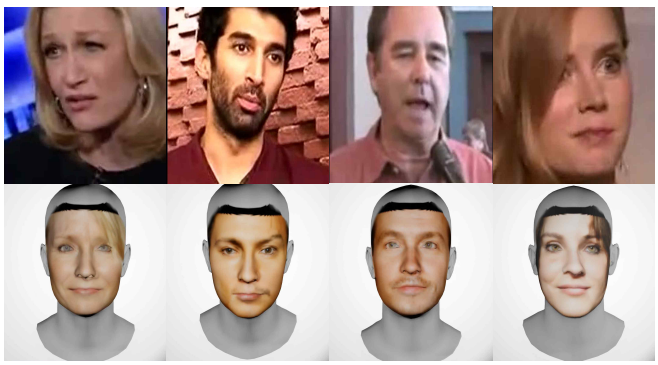

그림 5 (상) 발화자 얼굴 이미지 (하) 복원된 얼굴 메쉬

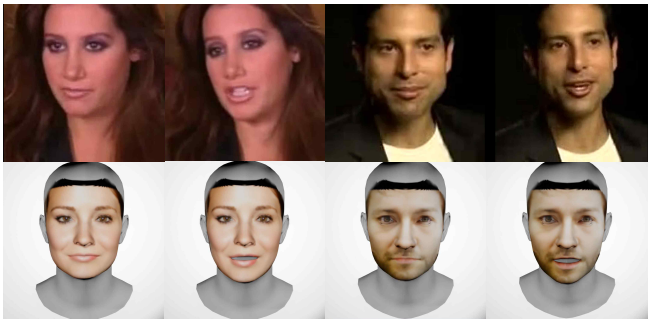

그림 6 (상) 발화자 얼굴 이미지 (하) 복원된 얼굴 메쉬

Ⅲ. 결론

본 논문에서는 짧은 음성 발화를 기반으로 3차원 얼굴 모델과 그 움직임을 복원하는 시스템을 제안하였다. 시스템을 통해 복원된 얼굴 메쉬가 발화자의 실제 얼굴과 공통적인 특성들을 가진다는 것을 확인하였다. 우리는 이 시스템이 가상현실 및 증강현실의 다양한 어플리케이션에서 활용될 것으로 기대한다.

ACKNOWLEDGMENT

이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00352, 설명 가능한 감성 경험 예측 모델 기반 콘텐츠 평가 기술 개발 및 상용화)

참 고 문 헌

[1] Fan, Yingruo, et al. "FaceFormer: Speech-Driven 3D Facial Animation with Transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Ephrat, Ariel, et al. "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation." arXiv preprint arXiv:1804.03619 (2018).

[3] Wen, Yandong, Bhiksha Raj, and Rita Singh. "Face reconstruction from voice using generative adversarial networks." Advances in neural information processing systems 32 (2019).

[4] Oh, Tae-Hyun, et al. "Speech2face: Learning the face behind a voice." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[5] Wu, Cho-Ying, Chin-Cheng Hsu, and Ulrich Neumann. "Cross-Modal Perceptionist: Can Face Geometry be Gleaned from Voices?." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[6] Li, Tianye, et al. "Learning a model of facial shape and expression from 4D scans." ACM Trans. Graph. 36.6 (2017): 194-1.

[7] Bulat, Adrian, and Georgios Tzimiropoulos. "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)." Proceedings of the IEEE International Conference on Computer Vision. 2017.