음성 기반 말하는 3D 얼굴 애니메이션 생성 연구 동향

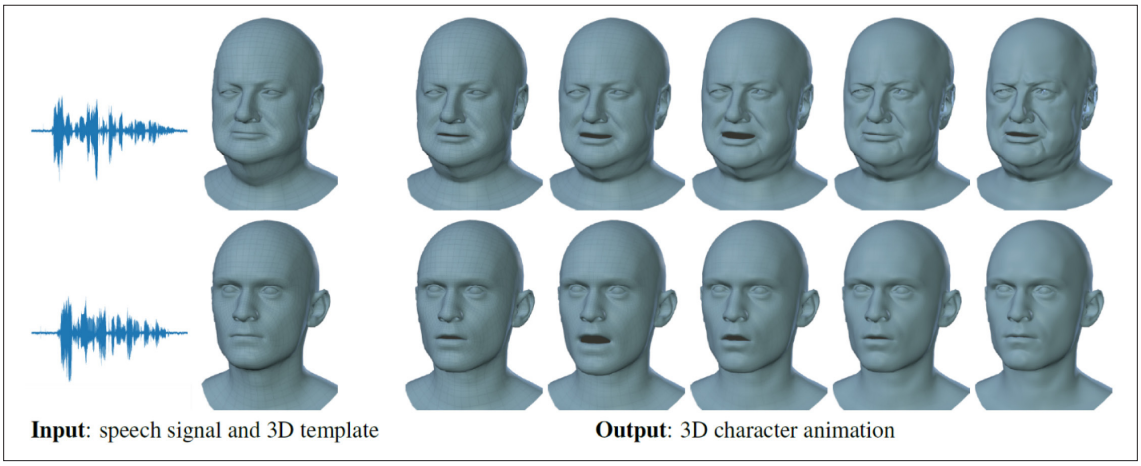
ㅁ 김학구 / 중앙대학교

요 약

음성 신호로부터 3D 얼굴 영상을 생성하는 연구는 최근 활발하게 연구되고 있는 기술로 게임, 영화 제작, 가상현실 분야 등에서 널리 활용될 수 있다. 음성 신호와 얼굴 움직임, 특히 입술의 움직임 간에는 높은 상관관계를 가지고 있기 때문에 충분한 데이터가 주어진 경우, 그 관계를 학습할 수 있다. 본 기고문에서는 음성 신호로부터 3D 얼굴 영상을 생성하는 기술(Speech-driven 3D Facial Animation)의 최근 동향과 향후 연구 방향에 대해서 논의한다. 이러한 방향성에서 대표적인 접근 방법론과 그 한계점을 소개하며, 이를 바탕으로 향후 연구 방향에 대해 논의하고자 한다.

I. 서 론

3D 디지털 인간은 증강현실(Augmented Reality)이나 가상현실(Virtual Reality)을 이용한 화상회의, 영화 제작이나 게임에서의 캐릭터 애니메이션 생성과 같이 몰입형 애플리케이션에서의 높은 활용도로 큰 관심을 받고 있다[1, 2]. 특히, 최근 신경 렌더링(Neural Rendering)[3,4]의 발전으로 인해 실제 인간의 외형과 움직임의 사실적인 합성에 있어서 엄청난 진전을 보이고 있다. 그러나 일부 애플리케이션에서는 텍스트나 오디오 입력만으로 3D 디지털 인간의 얼굴 아바타를 제어해야 할 필요가 있다. 예를 들어, 가상 화상회의(Telepresence) 중 얼굴이 가려져 추적할 수 없거나 웹캠이 없는 화상회의 환경 등에서는 오디오 입력만으로 작업해야 하는 경우가 있다. 이러한 상황을 대비하거나 보다 편의성 높은 애플리케이션 활용을 위해서 입력 음성 신호에서 3D 얼굴 애니메이션을 생성하는 기술이 요구된다(〈그림 1〉). 음성 신호 기반 3D 얼굴 애니메이션 생성 연구 환경에서는 입력으로 음성 신호와 3D 얼굴 템플릿 데이터가 메쉬 형태로 주어진다. 그 다음, 3D 메쉬 형태의 얼굴 템플릿의 정체성(identity)은 유



<그림 1> 음성 기반 말하는 3D 얼굴 애니메이션 생성 연구 개요도[1]

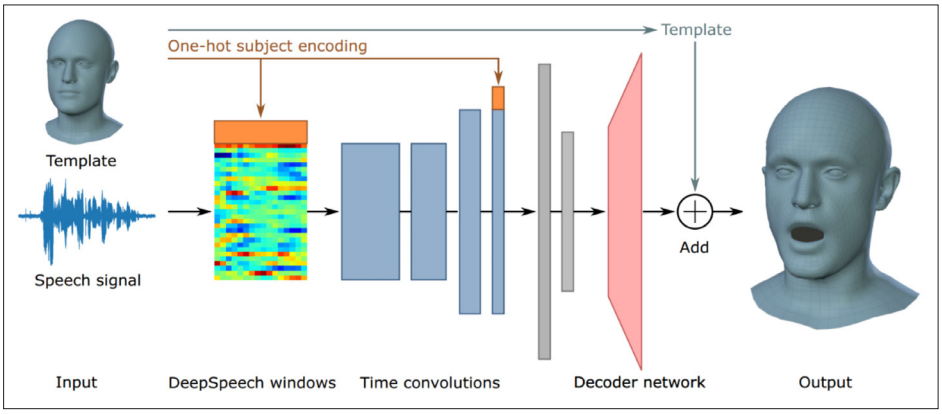
지하는 동시에 해당 음성 신호에 맞게 적절하게 표정과 입술을 움직이는 애니메이션 형태로 생성하는 것이 해당 기술의 목적이다.

최근 음성 기반 말하는 3D 얼굴 애니메이션 생성 연구는 충분한 데이터셋을 활용하여 음성 신호와 얼굴 움직임 간의 잠재적인 관계를 학습하는 방식으로 이루어진다[1,2,5,6,7,8]. 대부분의 최근 연구들은 고품질의 3D 얼굴 움직임 캡쳐 데이터셋에 대해 학습이 진행되며, DeepSpeech[9]나 wav2vec[10]과 같은 사전 학습된 음성 모델을 활용해 오디오 특징을 추출하는 구조를 포함하여 고품질의 얼굴 움직임을 생성하고자 한다. 그럼에도 불구하고 실제 사람과 같은 얼굴 움직임(표정, 입술)을 생성하는 것은 여전히 도전적이다. 특히, 모든 입력 음성 신호들로부터 하나의 3D 얼굴 생성 모델을 통해서 그에 맞는 다양한 사람들의 3D 얼굴 애니메이션을 생성하는 것은 본질적으로 불완전한 문제(ill-posed problem)이다. 결과적으로 이러한 모호성 때문에 많은 방법들이 시각적으로 과도하게 부드러운 애니메이션 결과를 야기하는 경향이 있다. 개인에 특화된 방법으로 상대적으로 개인별 적절한 3D 얼굴 움직임 생성 결과를 얻을 수도 있지만, 이러한 경우에는 일반적인 응용 애플리케이션으로써 확장성은 낮아진다[11,12]. 비교적 최근에 VOCA[1] 모델이 데이터셋과 함께 음성 신호로부터 얼굴 모션으로의 매핑 문제를 회귀 모델로 공식화함으로써 다양한 사람의 얼굴 애니메이션 생성에 대해 일반화 가능성을 보여줬다. 하지만 사람의 얼굴에는 입술과 같이 음성 신호와 밀접한 관계가 있는 영역이 있는가 하면, 코나 이마와 같이 음성 신호와 관련성이 적거나 거의 움직임이 없는 영역도 존재하기 때문에, 이러한 접근법은 음성 신호와 관련이 높은 입술이나 표정 움직임을 직접적으로 조절 및 생성하도록 함에 있어서 한계가 있다. 이러한 한계점을 해결하기 위해 Transformer 기반의 자동 회귀(Auto-regression) 방식으로 불확실성을 제거하는 방식[5]이나 음성 신호와 관련된 정보와 관련되지 않은 정보를 분리해서 처리하는 방식[7] 등 다양한 기법들이 제안되고 있다. 최근에는 단순히 음성 신호와 동기화된 입술 움직임 생성을 넘어서 같은 음성 신호가 주어지더라도 개인별 얼굴의 구조나 발화 특성 등에 따라 조금씩 다르게 개인화된 3D 얼굴 애니메이션 생성을 수행하는 기법들도 등장하고 있다[13,14].

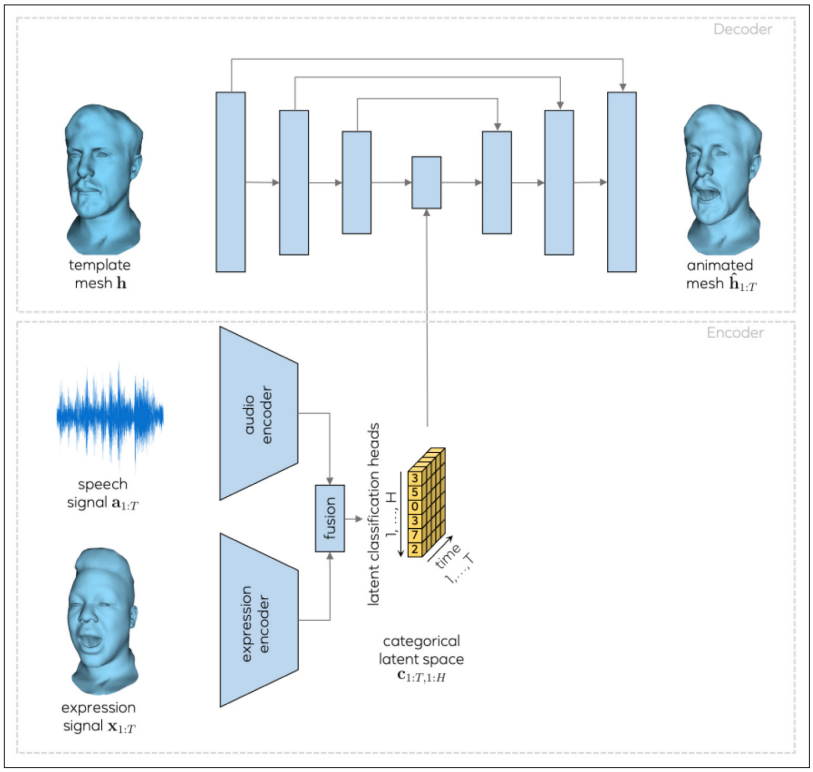
본고에서는 최근에 제안되고 있는 고품질 데이터셋을 학습하여 음성 신호로부터 말하는 3D 얼굴 애니메이션을 생성하는 방법에 대해 심도있게 살펴보고자 한다. II장에서는 다양한 심층신경망 구조와 학습 기법을 바탕으로 제안된 최신 기술 동향을 소개한다. III장에서는 주요 데이터셋과 최신 방법의 성능을 살펴보고, IV장에서 결론을 맺는다.

II. 음성 기반 말하는 3D 얼굴 애니메이션 생성 기술 동향

음성 신호로부터 3D 얼굴 애니메이션을 생성하는 연구는 매우 도전적인 기술로 인공지능 모델의 발전과 고품질 데이터셋의 등장과 함께 최근에 들어서 빠른 속도로 연구가 진행되고 있다. 해당 연구의 포문을 연 대표적인 기술인 VOCA[1]는 사전 학습된 DeepSpeech[9] 모델을 이용하여 음성 신호를 추출한 뒤, temporal convolutional neural networks(Temporal CNN)을 이용하여 시간 축



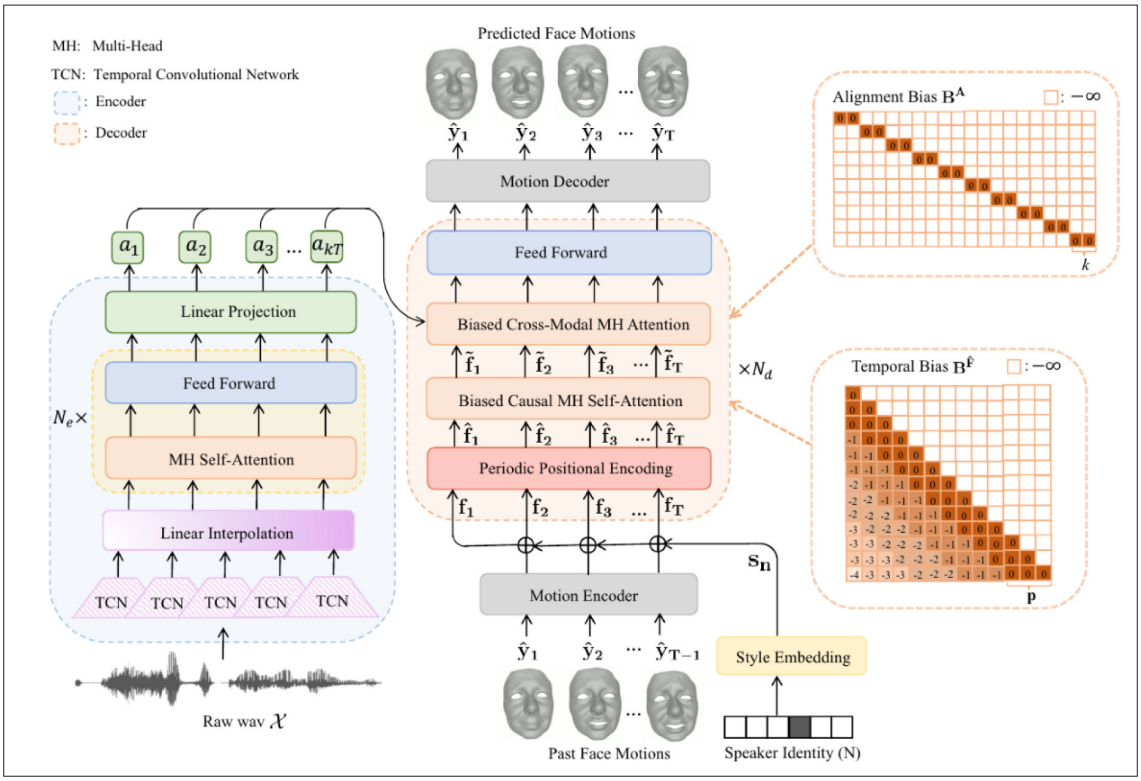
<그림 2> VOCA 기법의 네트워크 구조[1]



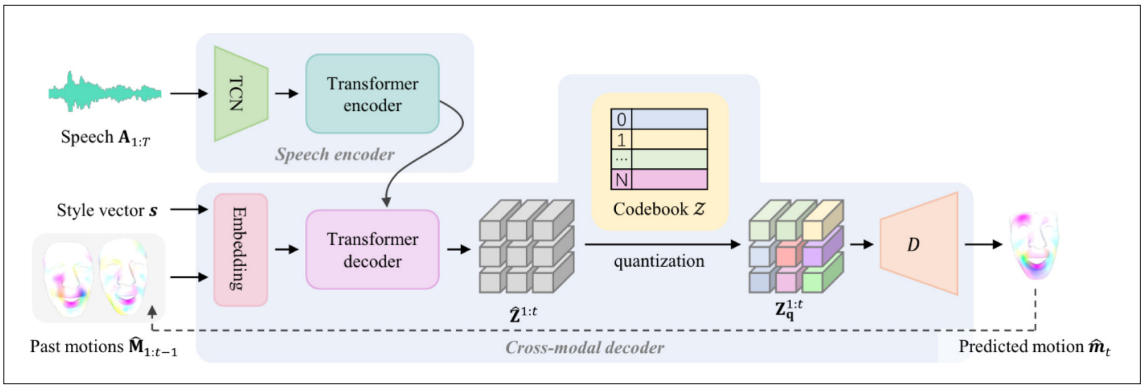
<그림 3> MeshTalk 기법의 전체 구조[7]

인코딩을 수행하고, 디코딩한 결과를 입력 3D 메쉬 형태의 얼굴 템플릿에 더하는 방식으로 말하는 얼굴 애니메이션을 생성한다(〈그림 2〉). MeshTalk[7]에서는 얼굴 영역 내 음성 신호와 관련이 높은 영역(예: 입술)과 그렇지 않은 영역(예: 눈썹)이 있다는 점에 주목하였다. 따라서, 음성 신호와 관련 있는 정보와 관련 없는 정보를 분리하는 categorical latent space를 학습하는 새로운 방식을 제안하였다(〈그림 3〉). 상대적으로 음성 신호와 관련이 적은 얼굴의 위쪽 영역은 주어진 입력 얼굴 템플릿 정보를 잘 활용하고, 반면에 입 주변 영역은 인코딩된 음성 신호로부터 잘 생성하도록 함으로써 성능 개선을 달성하였다. 하지만 이들은 모두 짧은 오디오 윈도우를 가지고 있어 음소(phoneme) 단위만을 학습하는 것에 집중한다. 그로 인해 실제로 문장 단위의 음성 신호 입력에 대해 때때로 부정확한 입술 움직임이 생성된다는 한계점을 지닌다.

음소 단위 오디오 인코딩으로 인한 생성 한계점을 극복하기 위해 등장한 모델이 Transformer 기반의 FaceFormer[5]이다(〈그림 4〉). FaceFormer[5]에서는 보다 긴 오디오 컨텍스트 정보 인코딩과 부족한 3D 오디오-메쉬 데이터 문제를 해결하고자 Transfomer를 활용한 자동 회귀(Auto-regression) 방식을 채택했다. 이렇게 함으로써 1) 보다 긴 오디오 컨텍스트 정보를 인코딩하는 것이 가능해짐으로써 얼굴 전체에 대해 보다 현실적인 말하는 얼굴 생성이 가능해졌고, 2) 자가학습(self-supervised) 기반의 사전학습된 음성 표현 모델을 활용하여 데이터 부족 문제를 다룰 수 있었으며, 3) 생성된 얼굴 움직임 정보를 계속적으로 고려함으로써 시간 축으로 안정적인 생성 결과를 획득할 수 있게 되었다. 하지만 여전히 해당 문제를 회귀 문제로 풀어나감으



<그림 4> FaceFormer 기법의 전체 구조도[5]

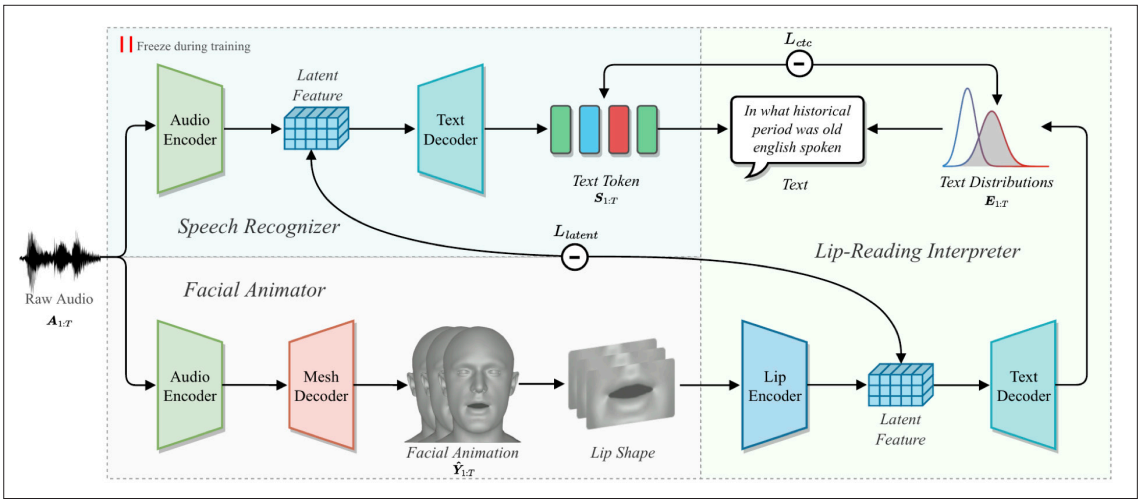


<그림 5> CodeTalker 기법의 전체 구조도[8]

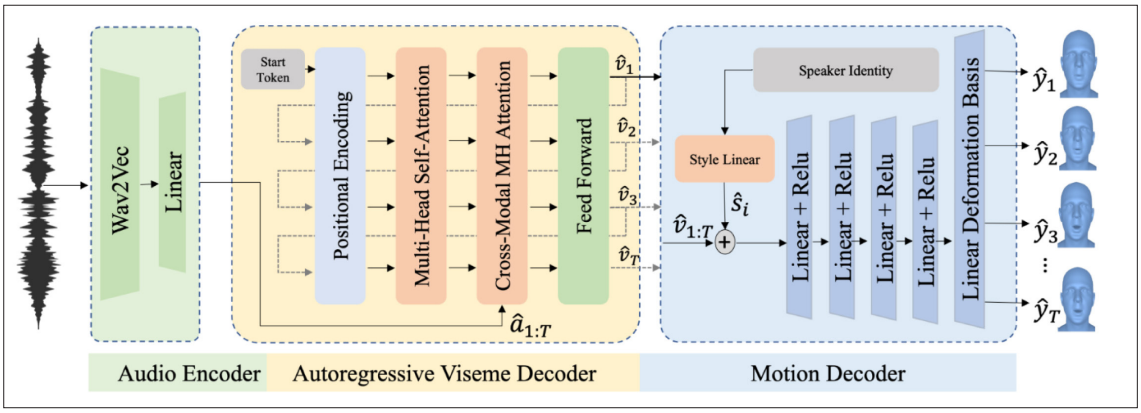
로써 말하는 움직임이 과도하게 부드러워지는 현상을 피할 수는 없었다.

2023년에 제안된 CodeTalker[8]에서는 음성 신호로부터 3D 얼굴 움직임을 생성하는 문제를 학습 가능한 이산화된 코드북이 있을 때, 그 유한한 공간에서의 코드-쿼리 문제로 해석하였다(〈그림 5〉). 코드북을 학습하기 위해서 vector-quantized autoencoder(VQ-VAE)가 사용되었으며, 실제 얼굴 움직임 데이터를 자가 복원(self-reconstruction)하는 방식으로 학습되었다. 구체적으로 입력 음성 신호가 들어오면 그로부터 움직임 특징 토큰을 예측하고, 이를 쿼리로 사용하여 사전 학습된 코드북에 의해 양자화되어 움직임 코드로 변환된다. 최종적으로 움직임 코드로부터 얼굴 움직임을 생성하게 되는데 이 과정을 순환하는 방식으로 말하는 얼굴 움직임을 예측한다.

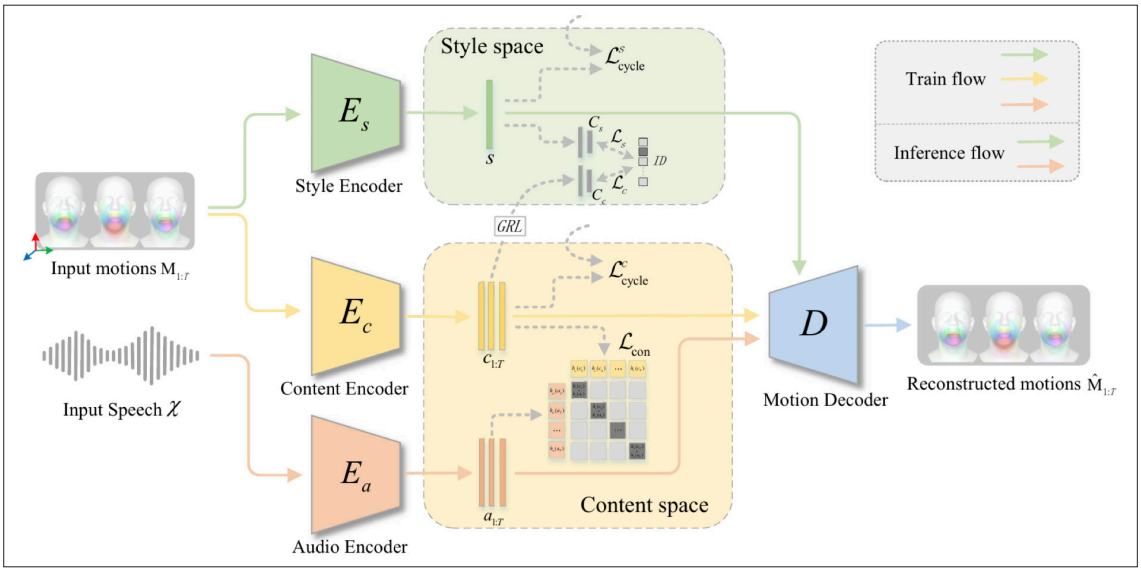
VOCA 모델의 등장 이후, 다양한 기법들이 제안되며 음성 기반 3D 얼굴 애니메이션 생성 연구가 발전하였으나, 해당 연구 주제의 특성이 특별히 반영되었거나 특별히 고안된 방법은 아니었다. 기본적인 U-Net 구조, Transformer 구조나 코드북 등 이미 다른 분야에서 많이 활용되고 있는 구조와 기법을 적용한 연구



<그림 6> SelfTalk 기법의 전체 개요도[15]



<그림 7> Imitator 기법의 딥 네트워크 구조[2]



<그림 8> Mimic 기법의 전체 구조도[13]

가 많았다.

SelfTalk[15]은 이러한 단순한 적용에서 벗어난 기법 중 하나로 라벨링된 데이터셋에 대한 의존성과 생성된 입술 움직임의 시각적 품질을 개선하고자 했다. SelfTalk은 크게 facial animator, speech recognizer, lip-reading interpreter의 3가지 모듈로 설계되었다. Facial animator를 통해 얼굴 애니메이션을 생성하고 이때, 생성된 입술 움직임에 대해 lip-reading을 수행했을 때 인식되는 문장과 음성 인식을 통해 오디오로부터 인식된 문장이 같아지도록 학습하는 구조이다(〈그림 6〉). 즉, 입력 데이터인 오디오 신호가 포함하고 있는 문자 정보와 오디오 신호로부터 생성된 입술 움직임이라는 시각 정보가 포함하고 있는 문자 정보는 같아야 한다는 개념을 설계한 것이다. SelfTalk[15]은 좋은 아이디어로 정확한 입술 움직임 생성을 달성하였지만, 기존 기법들과는 달리 주어진 얼굴마다의 정체성(identity)은 고려하지 않은

방법이다.

최근에 등장하고 있는 3D 얼굴 애니메이션 생성 연구는 화자의 얼굴 특징 및 발화 스타일까지 고려한 개인화된 3D 얼굴 애니메이션 생성 연구로 그 범위가 확장되고 있다. Imitator[2] 기법에서는 화자의 정체성 정보를 마지막 얼굴 움직임을 생성하는 디코더에서 결합하는 방식을 통해서 음성 신호로부터 전체적으로 공통적인 발음과 같은 비짐(viseme)을 디코딩하는 영역과 분리하였다(〈그림 7〉). 그렇게 함으로써 보다 정확한 입술 모양과 개별 얼굴 특징을 잘 생성하는 결과를 얻을 수 있었다. Mimic[13] 기법은 얼굴 움직임 안에 혼재되어 있는 화자별 발화 스타일과 음성 컨텐츠 정보를 명시적으로 분리하고자 시도했다(〈그림 8〉). Content contrastive loss를 이용하여 얼굴 정보 내에서 음성과 관련된 의미 정보를 담고 있는 부분과 발화 스타일과 같이 음성에 대한 의미 정보를 담고 있지 않은 부분을 구별함으로써 정확한 얼굴 애니메이션 생성이 가능하게 했다.

III. 성능 평가

1. 데이터셋

음성 신호 기반 3D 얼굴 애니메이션 생성 연구에 사용되는 데이터셋에는 크게 두 가지 종류가 있다. 첫 번째는 VOCA 모델을 제안한 저자들이 공개한 VOCASET 데이터이다[1]. VOCASET[1]은 12명의 화자로부터 약 3, 4초 길이의 480개 문장 시퀀스로 구성된 총 29분 길이의 4D 얼굴 스캔 및 오디오 데이터셋이다(https://voca.is.tue.mpg.de/). 두 번째 데이터셋은 BIWI 데이터이다[16]. BIWI[16]는 20명의 화자에 대해 15,000 이상의 이미지로 구성된 데이터로, 640x480 해상도의 RGB 이미지와 깊이 이미지가 쌍으로 매 프레임 구성되어 있다. 머리 각도는 yaw 축에 대해 ±75도, pitch 축에 대해 ±60도의 범위를 갖는 데이터이다(https://www.kaggle.com/datasets/kmader/biwi-kinect-head-pose-database).

2. 성능 비교

본 연구 결과의 정량적인 성능 평가를 위해서는 주로 face vertex error(FVE), lip vertex error(LVE), face/lip dynamics deviation(FDD/LDD), dynamic time wrapping(DTW), style cosine similarity(SCS) 등이 사용된다. 각각의 논문마다 학습과 평가에 사용하는 데이터셋과 성능 지표가 상이하여 공통된 실험 환경에서의 성능 비교를 하기에 어려운 점이 존재한다.

〈표 1〉은 VOCASET 데이터[1]에서 LVE와 FDD를 측정한 정량적 성능 비교 결과를 보여준다. 2019년도에 제안된 VOCA[1] 모델부터 최근에 제안된 SelfTalk[15]에 이르기까지 점진적으로 성능이 개선되는 경향성을 파악할 수 있다.

<표 1> VOCASET[1]에서의 다양한 기법 성능 비교

Methods	LVE↓ ($x10^{-5}$ mm)	FDD↓ ($x10^{-7}$ mm)
VOCA[1]	4.9245	4.8447
MeshTalk[7]	4.5441	5.2062
FaceFormer[5]	4.1090	4.6675
CodeTalker[8]	3.9445	4.5422
SelfTalk[15]	3.2238	4.0912

〈그림 9〉는 주어진 음성 신호로부터 각각의 방법으로 생성된 3D 얼굴 애니메이션과 ground-truth 이미지를 보여주고 있다. 어려운 발음의 경우, VOCA와 MeshTalk의 경우, 제대로 입술 움직임이 생성되지 않는 반면, FaceFormer와 Imitator는 잘 생성하는 것을 볼 수 있다. 특히, 가장 최신 방법 중 하나인 Imitator[2] 기법이 보다 정확한 입술 모양을 생성한 결과를 확인할 수 있으며, 앞에서 소개한 각 방법들의 한계점과 일관된 결과가 정성적인 결과로도 드러나는 것을 관찰할 수 있다.

<그림 9> 다양한 기법들의 3D 얼굴 애니메이션 생성 결과 비교[2](Ours: Imitator[2])

IV. 결 론

본고에서는 최근 컴퓨터 비전 분야에서 활발하게 연구되고 있는 연구 중 하나인 음성 신호 기반 말하는 3D 얼굴 애니메이션 생성 기법에 대한 최근 연구 동향에 대해 논의하였다. 음성 신호만을 가지고 어떤 사람의 표정과 입술 움직임을 생성하는 기술은 최근 주목받고 있는 가상현실/메타버스 플랫폼에서의 높은 활용도로 큰 주목을 받고 있다. 이전까지는 데이터셋의 부족과 음성 신호와 3D 얼굴 기하 정보 간 상관관계 분석의 어려움 등으로 많은 연구가 진행되지 못했다. 최근 들어 고품질 데이터셋의 등장과 각 도메인에서 심층신경망의 발전에 힘입어 해당 기술이 빠르게 발전하고 있다. 앞으로는 음성 신호로부터 단순히 일반적인 입술의 움직임을 생성하는 것을 넘어서 사람마다 가지고 있는 표정이나 입술 움직임의 특징까지 고려하여 개인화된 3D 얼굴 표정/감정/입술 애니메이션을 생성하는 연구로 확장될 것으로 기대된다. 더욱이 해당 기술은 영화나 게임과 같은 문화 콘텐츠 산업과 가상현실에서의 높은 잠재력과 활용성을 가지고 있으므로 앞으로 지속적인 최신 연구 동향에 대한 관심과 논의가 계속 필요할 것이다.

참 고 문 헌

[1] Cudeiro D, Bolkart T, Laidlaw C, Ranjan A, Black MJ. Capture, learning, and synthesis of 3D speaking styles. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 (pp. 10101-10111).
[2] Thambiraja B, Habibie I, Aliakbarian S, Cosker D, Theobalt C, Thies J. Imitator: Personalized speech-driven 3d facial animation. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2023 (pp. 20621-20631).
[3] Tewari A, Thies J, Mildenhall B, Srinivasan P, Tretschk E, Yifan W, Lassner C, Sitzmann V, Martin-Brualla R, Lombardi S, Simon T. Advances in neural rendering. In Computer Graphics Forum 2022 May (Vol. 41, No. 2, pp. 703-735).
[4] Tewari A, Fried O, Thies J, Sitzmann V, Lombardi S, Sunkavalli K, Martin-Brualla R, Simon T, Saragih J, Nieβner M, Pandey R. State of the art on neural rendering. In Computer Graphics Forum 2020 May (Vol. 39, No. 2, pp. 701-727).
[5] Fan Y, Lin Z, Saito J, Wang W, Komura T. Faceformer: Speech-driven 3d facial animation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 18770-18780).
[6] Karras T, Aila T, Laine S, Herva A, Lehtinen J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (ToG). 2017 Jul 20;36(4):1-2.
[7] Richard A, Zollhöfer M, Wen Y, De la Torre F, Sheikh Y. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021 (pp. 1173-1182).
[8] Xing J, Xia M, Zhang Y, Cun X, Wang J, Wong TT. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 (pp. 12780-12790).
[9] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, Prenger R, Satheesh S, Sengupta S, Coates A, Ng AY. Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567. 2014 Dec 17.
[10] Schneider S, Baevski A, Collobert R, Auli M. wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862. 2019 Apr 11.
[11] Karras T, Aila T, Laine S, Herva A, Lehtinen J. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (ToG). 2017 Jul 20;36(4):1-2.
[12] Richard A, Lea C, Ma S, Gall J, De la Torre F, Sheikh Y. Audio-and gaze-driven facial animation of codec avatars. In Proceedings of the IEEE/CVF winter conference on applications of computer vision 2021 (pp. 41-50).
[13] Fu H, Wang Z, Gong K, Wang K, Chen T, Li H, Zeng H, Kang W. Mimic: Speaking Style Disentanglement for Speech-Driven 3D Facial Animation. In Proceedings of the AAAI Conference on Artificial Intelligence 2024 Mar 24 (Vol. 38, No. 2, pp. 1770-1777).
[14] Wu H, Zhou S, Jia J, Xing J, Wen Q, Wen X. Speech-driven 3d face animation with composite and regional facial movements. In Proceedings of the 31st ACM International Conference on Multimedia 2023 Oct 26 (pp. 6822-6830).
[15] Peng Z, Luo Y, Shi Y, Xu H, Zhu X, Liu H, He J, Fan Z. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In Proceedings of the 31st ACM International Conference on Multimedia 2023 Oct 26 (pp. 5292-5301).
[16] Fanelli G, Gall J, Romsdorfer H, Weise T, Van Gool L. A 3-d audio-visual corpus of affective communication. IEEE Transactions on Multimedia. 2010 Sep 13;12(6):591-8.

저 자 소 개

김 학 구

- 2012년 : 인하대학교 전자공학과 학사
- 2014년 : 인하대학교 전자공학과 석사
- 2019년 : 한국과학기술원 전기 및 전자공학부 박사
- 2019년 ~ 2021년 : 스위스 로잔연방공과대학 (EPFL) 박사후연구원
- 2021년 ~ 현재 : 중앙대학교 첨단영상대학원 조교수
- 주관심분야 : 3D/VR, 3D 생성모델, 멀티모달