ISSN 2508-2817 KiDRS Korea Institute of Design Research Society 통권 35호, 2025. Vol.10 No.2

3D 디지털 휴먼의 자연스러운 한국어 발화를 위한 AI 학습용 데이터세트 구축 사례

A Case Study on the Construction of an AI Training Dataset for Natural Korean Speech Generation by 3D Digital Humans

주 저 자 : 이 솔 (Lee, Sol) 오모션 주식회사 연구원
s.l@omotion.co.kr

공 동 저 자 : 이학범 (Lee, Hak Bum) 광운대학교 전자재료공학과 석사과정
hblee@kw.ac.kr

공 동 저 자 : 김진겸 (Kim, Jin Kyum) 광운대학교 전자재료공학과 박사과정
jkkim@kw.ac.kr

교 신 저 자 : 오문석 (Oh, Moon Seok) 광운대학교 미디어커뮤니케이션학부 교수
motion@kw.ac.kr

교 신 저 자 : 서영호 (Seo, Young Ho) 광운대학교 전자재료공학과 교수
yhseo@kw.ac.kr

https://doi.org/10.46248/kidrs.2025.2.122

접수일 2025. 05. 31. / 심사완료일 2025. 06. 02. / 게재확정일 2025. 06. 09. / 게재일 2025. 6. 30.
이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의지원(No. RS-2025-02216275, 사실적 움직임 생성·재현이 가능한 디지털 휴먼 기술)과 2023년도 정부(중소벤처기업부)의 재원으로 중소기업기술정보진흥원의 지원을 받아 수행된 연구임(No. 00264741, 비전기반의 모션캡쳐를 이용한 실시간 버추얼 스트리밍 솔루션 개발)

Abstract

Recently, the demand for 3D digital humans—capable of being used without spatial or temporal constraints—has been increasing in broadcasting and content production. However, there is a lack of 3D facial data tailored for natural Korean speech generation. This study aims to construct a Korean speech-driven 3D facial dataset for digital human creation and to develop a generative AI model that produces realistic facial animations. A total of over 570,000 high-resolution 3D meshes were built through multi-view capture of 5,000 carefully designed sentences that reflect articulatory features and phoneme distributions. Based on this dataset, a transformer-based CodeTalker model was trained to generate synchronized and lifelike 3D facial animations from speech input. This research establishes a foundation for Korean-specific digital human development and contributes to various domains such as the metaverse, broadcasting, and public services through the practical application of generative AI technologies.

Keyword

Speech-driven Facial Animation(음성 기반 얼굴 애니메이션), Korean Language Synthesis(한국어 합성), Multimodal Dataset Construction(멀티모달 데이터셋 구축)

요약

최근 방송 및 콘텐츠 제작 분야에서 시공간 제약 없이 활용 가능한 3D 디지털 휴먼의 수요가 증가하고 있으나, 자연스러운 한국어 발화를 위한 3D 얼굴 데이터는 부족한 실정이다. 본 연구는 3D 디지털 휴먼 제작을 위한 한국어 음성 기반 발화 얼굴 데이터를 구축하고, 이를 기반으로 생성형 AI 모델을 학습하여 현실감 있는 얼굴 애니메이션을 생성하는 것을 목표로 한다. 조음 특성과 음운 분포를 고려해 설계한 5,000문장 대본과 다시점 촬영을 통해 총 570,000개 이상의 고정밀 3D 메쉬 데이터를 구축하였으며, 트랜스포머 기반의 CodeTalker 모델을 학습해 음성과 표정 간의 정합성이 높은 3D 애니메이션을 구현하였다. 본 연구는 한국어 특화 디지털 휴먼 제작 기반을 마련하고, 생성형 AI 기술의 실용화를 통해 메타버스, 방송, 공공서비스 등 다양한 분야에 기여할 수 있다.

목차

1. 서론

1-1. 연구의 배경 및 중요성

최근 가상인간 기술은 빠르게 발전하며 다양한 산업 분야에서 활용되고 있다. 가상 인플루언서, 디지털 휴먼, 가상 캐릭터 등은 엔터테인먼트, 마케팅, 교육, 고객 서비스 등 여러 분야에서 중요한 역할을 맡고 있다. 이러한 기술은 사용자와의 상호작용을 보다 몰입감 있게 만들어주며, 가상세계와 현실을 연결하는 중요한 가교 역할을 한다. 특히 가상인간 애니메이션 기술은 현실감 있는 표현을 통해 가상 인물의 감정 및 행동을 사실적으로 구현할 수 있어 필수적이다. 이로 인해 가상인간 애니메이션 기술의 발전은 다양한 분야에서 더욱 중요한 기술로 자리 잡고 있으며, 이에 따른 연구와 기술 개발의 필요성이 더욱 커지고 있다.[1)]

그러나 가상인간 기술의 발전에는 언어와 관련된 한계가 존재한다. 특히, 한국어는 발음과 억양에서 다른 언어들과 구별되는 특징을 지니고 있기 때문에, 한국어를 위한 발화 및 표정 데이터셋은 상대적으로 부족하다.[2)] 이에 따라 한국어 기반의 3D 얼굴 애니메이션 생성 연구가 필수적이다. 한국어에 특화된 데이터셋이 부족한 상황에서, 기존의 다국어 데이터셋을 그대로 적용하는 데에는 정확성에 한계가 존재한다. 따라서 한국어에 최적화된 데이터와 기술 개발이 중요한 시점에 도달하였다.

또한, 3D 처리 기술의 발전과 컴퓨팅 성능의 향상은 2D 데이터에서 3D 데이터로의 진화를 가능하게 했다.[3)] 과거 2D 기반의 모델링 기술은 주로 정적이고 제한적인 범위에서 활용되었지만, 최근에는 3D 모델링과 애니메이션이 가능해지면서, 훨씬 더 사실적이고 몰입감 있는 가상 인물 및 환경을 창조할 수 있게 되었다. 3D 처리 기술의 발전은 가상 인간 및 얼굴 애니메이션 분야에서도 중요한 전환점을 맞이했으며, 이를 활용한 다양한 연구가 이루어지고 있다. 이러한 기술 발전은 가상인간이 단순한 시각적 표현을 넘어 실시간 상호작용이 가능한 디지털 휴먼으로 발전할 수 있는 가능성을 열어주었다.

2D 기반의 학습 모델들은 많은 시간과 비용을 요구한다.[4)] 특히, 2D 데이터셋은 정확한 얼굴 표정 생성 및 발화 동기화의 한계를 가지며, 이를 극복하기 위한 추가적인 비용과 시간이 소모된다. 매개변수화된 3D 모델은 이러한 2D 모델의 한계를 해결할 수 있는 가능성을 제공하며, 더 적은 자원으로 자연스러운 얼굴 애니메이션을 생성할 수 있다. 3D 모델을 활용하면 더 정교한 표정 및 발화 연동이 가능해지고, 실시간으로 보다 자연스러운 반응을 구현할 수 있다. 따라서 3D 기반의 연구는 앞으로의 기술 발전에 있어 매우 중요한 역할을 하게 될 것이다.

현재 대부분의 3D 얼굴 모델링은 고정밀 메쉬를 사용하지 않아 얼굴 표정 변화나 세밀한 디테일을 정확하게 표현하기 어려운 한계가 있다.[5)] 3D 메쉬는 얼굴 애니메이션의 정밀도에 큰 영향을 미치며, 이를 개선하는 것은 보다 사실적인 디지털 휴먼 구현의 핵심이다. 고정밀 메쉬를 적용하여 얼굴의 미세한 움직임까지 정교하게 표현할 수 있는 기술이 필요하며, 이는 가상인간의 현실감과 몰입감을 높이는 중요한 요소로 작용한다. 따라서, 고해상도 메쉬 생성 기술의 발전은 3D 얼굴 애니메이션 연구에서 중요한 연구 과제가 된다.

1-2. 선행 연구

한국어 음성 기반 3D 얼굴 데이터셋에 관한 연구는 크게 네 가지 분야로 나눌 수 있다: 1) 3D 얼굴 데이터셋 연구, 2) 음성 기반 3D 얼굴 애니메이션 연구, 3) 음성 기반 전신 제스쳐 애니메이션 연구, 4) 한국어(음성, 텍스트) 기반 3D 얼굴 애니메이션 연구. 이들 연구는 각각 3D 얼굴 애니메이션과 음성의 동기화를 다루며, 특히 한국어 음성 기반 3D 얼굴 데이터셋 제작과 학습에 대한 이해를 돕는 중요한 이론적 기반을 제공한다. 이러한 선행 연구들은 본 연구의 진행에 필수적인 기술적 및 이론적 토대를 마련하며, 음성과 3D 얼굴 애니메이션의 통합에 있어 중요한 참고자료로 활용될 수 있다.

---

1) 오문석, 한규훈, and 서영호. "메타버스를 위한 디지털 휴먼과 메타휴먼의 제작기법 분석 연구." 한국디자인리서치학회 6.3 (2021): 133-142.

2) 강석찬, and 김동주. "딥러닝을 활용한 한국어 스피치 애니메이션 생성에 관한 고찰." 정보처리학회논문지. 소프트웨어 및 데이터 공학 12.10 (2023): 461-470.

3) 김현주, et al. "고해상도 3D 데이터 생성 기술 분석 및 연구 동향." [ETRI] 전자통신동향분석 37.3 (2022): 0-0.

4) https://pixune.com/blog/2d-animation-vs-3d-animation-cost/

5) Ming, Xin, et al. "High-Quality Mesh Blendshape Generation from Face Videos via Neural Inverse Rendering." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

1-2-1. 3D 얼굴 데이터셋 연구

3D 얼굴 애니메이션 연구에서는 다양한 방식으로 얼굴 형상을 수집하거나 재구성한 데이터셋이 활용되고 있다. 첫째, 다시점 촬영과 3D 재구성 기반 데이터셋으로는 VOCASET[6]과 Multiface[7]가 대표적이다. VOCASET은 12명의 피험자가 다양한 문장을 발화하는 장면을 다각도에서 촬영한 후, FLAME 모델 기반으로 정렬된 고해상도 3D 메쉬를 구성한 데이터셋이다. Multiface는 9대의 RGB 카메라를 활용하여 조명과 시선, 표정이 다양한 상황에서 다각도 영상을 수집한 후, 멀티뷰 스테레오 기법으로 프레임 단위의 정밀한 3D 얼굴 메쉬를 생성하였다.

둘째, 3D 스캐너를 활용한 정밀 스캔 기반 데이터셋으로는 BIWI[8]가 대표적이다. 이 데이터셋은 무향실에서 고정된 조명과 고해상도 3D 스캐너를 사용하여 수집되었으며, 피험자가 감정 유도 영상을 시청한 후 감정을 담아 문장을 발화하는 장면을 3D 메쉬로 촬영하였다. BIWI는 초당 25프레임으로 정밀한 얼굴 정점(23,390개)을 기록하여, 시계열 기반의 감정 표현 분석에도 활용도가 높다.

마지막으로, 단일 영상 기반의 FLAME 또는 3DMM 추정 방식을 사용하는 데이터셋으로는 S3DFM[9], MultiTalk[10], LaughTalk[11], 3D-RAVDESS[12] 등이 있다. 이들은 2D 영상과 음성을 동시에 수집한 후, 프레임 단위로 3DMM 또는 FLAME 모델 피팅을 통해 표현 파라미터를 예측하여 pseudo-3D 데이터를 구축한다. 이러한 방식은 장비 부담이 적고 대규모 데이터 수집에 유리하나, 실제 3D 측정에 비해 정밀도가 다소 낮을 수 있다.

[표 1] 3D 얼굴 데이터셋 구축 방식에 따른 분류

구축 방식	데이터셋
다시점 촬영 및 3D 재구성	VOCASET, Multiface
3D 스캐너	BIWI
단일 영상 기반 3D 추정	S3DFM, MultiTalk, LaughTalk, 3D-RAVDESS

1-2-2. 음성 기반 3D 얼굴 애니메이션 연구

음성을 입력으로 받아 3D 얼굴 애니메이션을 생성하는 기술은 최근 다양한 접근 방식으로 발전해왔다. 그중에서도 모션 임베딩 기반 방식은 음성 신호에서 고수준의 시간-공간적 움직임 정보를 추출해 보다 일관된 모션 시퀀스를 생성하는 데 집중한다. KMTalk[13]은 음소 정렬 기반의 Key Motion Embedding을 통해 발화 구조를 정교하게 반영하며, CodeTalker[14]는 실제 3D 모션 데이터를 압축한 이산 코드북을 통해 사실적인 입술 움직임을 복원한다. 이러한 접근은 학습 효율성과 현실감을 동시에 확보하는 데 유리하다.

다른 한편으로, 최근에 감정 표현을 포함한 애니메이션 생성도 활발히 연구되고 있다. 3DFacePolicy[15]는 디퓨전 정책 학습을 통해 감정이 담긴 얼굴 정점 궤적을 연속적으로 예측하고, DEITalk[16]은 MetaHuman 기반 감정 데이터셋을 활용해 감정 강도

6) Cudeiro, Daniel, et al. "Capture, learning, and synthesis of 3D speaking styles." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

7) Wuu, Cheng-hsin, et al. "Multiface: A dataset for neural face rendering." arXiv preprint arXiv:2207.11243 (2022).

8) Fanelli, Gabriele, et al. "A 3-d audio-visual corpus of affective communication." IEEE Transactions on Multimedia 12.6 (2010): 591-598.

9) Zhang, Jie, and Robert B. Fisher. "3d visual passcode: Speech-driven 3d facial dynamics for behaviometrics." Signal processing 160 (2019): 164-177.

10) Sung-Bin, Kim, et al. "MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset." arXiv preprint arXiv:2406.14272 (2024).

11) Sung-Bin, Kim, et al. "Laughtalk: Expressive 3d talking head generation with laughter." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.

12) Liu, Chang, et al. "EmoFace: Audio-driven emotional 3D face animation." 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, 2024.

13) Xu, Zhihao, et al. "KMTalk: Speech-Driven 3D Facial Animation with Key Motion Embedding." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.

14) Xing, Jinbo, et al. "Codetalker: Speech-driven 3d facial animation with discrete motion prior." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

15) Sha, Xuanmeng, et al. "3DFacePolicy: Speech-Driven 3D Facial Animation with Diffusion Policy." arXiv preprint arXiv:2409.10848 (2024).

의 시간적 흐름을 모델링한다. EmoTalk[17]은 감정과 내용을 음성에서 분리해 다양한 감정 스타일을 적용할 수 있는 구조를 제안함으로써, 표정 표현의 다양성과 사실성을 향상시킨다. 이들은 단순한 입모양 예측을 넘어 감정 이입 가능한 애니메이션 생성에 중점을 두고 있다.

한편, DiffPoseTalk[18], GLDiTalker[19], SelfTalk[20]은 기존 범주에 쉽게 속하지 않지만, 각각 독자적인 방법론을 통해 새로운 접근을 시도한다. DiffPoseTalk은 참조 영상에서 스타일을 추출하고 디퓨전 모델을 통해 머리 움직임과 얼굴 포즈까지 포함한 다채로운 애니메이션을 생성한다. GLDiTalker는 그래프 기반 잠재공간에서 모달리티 불일치를 완화하며, 입술 싱크와 모션 다양성 모두를 향상시키는 구조를 채택한다. SelfTalk은 라벨이 없는 상태에서도 자기 지도 학습을 수행할 수 있도록 설계되었으며, 음성으로부터 생성된 3D 얼굴 메쉬에서 추출한 텍스트 코드와, 음성으로부터 직접 추출한 텍스트 코드를 비교함으로써 음성과 얼굴 애니메이션 간의 의미론적 정합성을 학습한다. 이를 통해 명시적 라벨 없이도 정밀한 입술 움직임을 생성할 수 있는 모델 구조를 제안한다.

이처럼 음성 기반 3D 얼굴 애니메이션 연구는 음성 구조 학습, 감정 표현, 스타일 분리, 자기지도 학습 등 다양한 방향에서 진화를 거듭하고 있으며, 각각의 방식은 실제 애니메이션 응용 목적에 따라 선택적으로 활용될 수 있다.

16) Shen, Kang, et al. "DEITalk: Speech-Driven 3D Facial Animation with Dynamic Emotional Intensity Modeling." Proceedings of the 32nd ACM International Conference on Multimedia. 2024.

17) Peng, Ziqiao, et al. "Emotalk: Speech-driven emotional disentanglement for 3d face animation." Proceedings of the IEEE/CVF international conference on computer vision. 2023.

18) Sun, Zhiyao, et al. "Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models." ACM Transactions on Graphics (TOG) 43.4 (2024): 1-9.

19) Lin, Yihong, et al. "Glditalker: Speech-driven 3d facial animation with graph latent diffusion transformer." arXiv preprint arXiv:2408.01826 (2024).

20) Peng, Ziqiao, et al. "Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces." Proceedings of the 31st ACM International Conference on Multimedia. 2023.

[표 2] 3D 얼굴 애니메이션 생성 연구의 접근 방식별 분류

방식	3D 얼굴 애니메이션 생성 연구
모션 임베딩	KMTalk, CodeTalker
감정 표현	3DFacePolicy, DEITalk, EmoTalk
디퓨전 모델	DiffPoseTalk
그래프 기반 모델	GLDiTalker
자기지도학습	SelfTalk

1-2-3. 한국어(음성, 텍스트) 기반 3D 얼굴 애니메이션 연구

한국어 스피치 애니메이션은 언어 고유의 음운론적 특성과 조음 체계로 인해 영어 기반의 기존 연구 결과를 그대로 적용하기 어렵다는 한계를 갖는다. 이에 따라, 2020년에는 한국어 자소 구조와 동시조음 현상을 반영한 규칙 기반 모델이 제안되었다.[21] 이 모델은 자모 단위의 viseme 정의와 위치 기반의 예외 처리 규칙을 적용해, 자음·모음 조합에 따른 입술 및 혀의 움직임을 정교하게 조절하였다. 특히 운율 요소(발화 속도, 세기, 길이 등)를 반영해 입모양의 타이밍과 쉐입 보간을 개선함으로써, 보다 자연스러운 애니메이션을 생성하였다.

한편 2023년에는 지도학습 기반의 딥러닝 모델이 최초로 한국어 스피치 애니메이션에 적용되었다.[22] 이 연구는 음성 인식 모듈과 표정 코딩 모듈을 분리하여 구성한 파이프라인 구조를 통해, 음성을 grapheme 시퀀스로 변환한 뒤 blendshape 애니메이션으로 생성하는 방식을 채택하였다. 실험 결과, 한국어 특화 음성 인식기를 적용했을 때 영어 기반 인식기보다 훨씬 더 자연스러운 애니메이션을 생성할 수 있음을 유저 스터디를 통해 입증하였다. 두 연구는 각각 규칙 기반과 데이터 기반이라는 접근 차이가 있으나, 모두 한국어 특성에 맞춘 음운 정보의 중요성을 강조하며 한국어 스피치 애니메이션의 정밀도

21) 장민정, 정선진, and 노준용. "한국어 동시조음 모델에 기반한 스피치 애니메이션 생성." 한국컴퓨터그래픽스학회논문지 26.3 (2020): 49-59.

22) 강석찬, and 김동주. "딥러닝을 활용한 한국어 스피치 애니메이션 생성에 관한 고찰." 정보처리학회논문지. 소프트웨어 및 데이터 공학 12.10 (2023): 461-470.

향상에 기여하였다.

1-3. 선행 연구의 한계 및 연구 목적

앞서 검토한 선행 연구들은 한국어 음성 기반 3D 발화 얼굴 데이터셋 구축에 관한 중요한 이론적, 기술적 토대를 제공하고 있으나, 한국어 특성에 최적화된 3D 얼굴 애니메이션 데이터셋 개발에는 몇 가지 주요한 한계점이 존재한다. 특히, 기존의 연구들은 주로 영어 또는 다른 언어를 기반으로 한 데이터셋을 사용하였으며, 한국어 발음과 억양 특성을 정확하게 반영한 고해상도 3D 얼굴 데이터셋과 애니메이션 기술의 발전에 대한 연구는 부족한 실정이다. 이로 인해 본 연구에서는 한국어 음성에 최적화된 3D 얼굴 애니메이션 생성 기술을 개발하고자 한다.

○ 대부분의 기존 데이터셋이 영어 기반으로 구축되어 한국어의 고유한 음운 특성과의 불일치를 초래함

○ 조음 특성을 고려한 문장 단위 음성 데이터 구성 및 분포 설계에 대한 체계적 기준 부재

○ 실제 3D 측정 없이 2D 추정 기반으로 구축된 pseudo-3D 데이터셋의 정밀도와 표현력 한계

○ 감정이나 억양의 뉘앙스를 반영한 한국어 애니메이션 연구가 아직 초기 단계에 머무름

앞서 분석한 선행 연구의 한계점을 극복하고 한국어 음성 기반 3D 발화 얼굴 데이터셋 구축 방법을 개발하기 위해, 본 논문은 아래와 같은 네 가지 목적을 가진다.

○ 문장 구성 시 조음 특성, 음운 분포, 문장 길이 등을 고려한 체계적 녹음 대본 설계 및 구축

○ 한국어 음성의 자소 단위 조음 특성을 반영한 고정밀 3D 발화 얼굴 데이터셋 구축

○ 고해상도 메쉬 기반의 3D 얼굴 표현으로 사실적인 입 모양 및 표정 데이터 확보

○ 한국어 음성과 3D 얼굴 애니메이션 간의 정합성을 학습하는 데이터 기반 모델 설계

이러한 연구 목적을 달성함으로써, 본 논문은 한국어 음성 기반의 고정밀 3D 발화 얼굴 데이터셋을 구축하고자 한다. 이는 가상인간 기술이 급격히 발전하고 다양한 분야에서 활용되는 현 시점에서, AI 및 컴퓨터 비전 연구자들과 VR/AR 및 게임 산업 관계자들에게 한국어 발화 기반의 3D 얼굴 애니메이션 학습 및 활용을 위한 중요한 데이터셋을 제공하며, 한국어 음성을 기반으로 한 가상인간 개발의 정확도와 현실감을 향상시키는 데 기여할 것이다.

2. 문장의 구성

본 연구의 목적은 한국어 발화 얼굴 생성을 위한 생성형 인공지능(AI) 개발에 필요한 음성 기반 발화 얼굴 데이터베이스 구축을 위한 녹음용 대본 문장의 선정 기준과 절차를 제시하는 데 있다. 특히, 제안된 5,000문장 규모의 녹음 대본이 어떠한 기준에 따라 선별되고 정제되었는지를 구체적으로 기술하고자 한다. 본 절에서는 연구팀이 대규모 한국어 말뭉치를 바탕으로 문장의 길이, 음운 분포, 특정 음운의 문장 내 위치 등을 고려하여 문장을 추출하고 정련한 절차를 상세히 설명한다.

본 데이터 구축은 생성형 AI 학습을 위한 음성·영상 통합 데이터베이스 구축이라는 실용적 목적에 따라 수행되었으며, 요구된 조건은 발화 시간 기준 4초에서 5초 정도에 해당하는 문장 5,000개였다. 이에 따라 문장 선정 과정에서는 조음 동작에 영향을 미칠 수 있는 음운을 고려하고, 이들이 문장 내 다양한 위치에 분포하도록 하는 것을 핵심 원칙으로 설정하였다. 특히, 입술과 턱 등 발화 얼굴에 시각적으로 큰 영향을 미치는 조음 동작이 문장의 초두 및 중간 위치에 나타날 수 있도록 문장을 구성하였다.

문장 선정 시 주요하게 고려된 기준은 다음과 같다.

① 발화 얼굴에 영향을 미칠 수 있는 조음 동작을 반영할 것

② 특정 조음 동작이 문장의 시작 및 중간 위치에 고르게 분포하도록 할 것

③ 실제 언중이 사용하는 자연스러운 문장을 포함할 것

④ 구어체와 문어체 문장을 균형 있게 포함할 것

⑤ 문장 길이는 발화 시간 기준 4초~5초에 해당하도록 조정할 것

기준 ①과 ②는 다양한 조음 동작이 문장 내에 자연스럽게 반영되도록 하기 위한 것이다. 특히 기준 ①은 입술과 턱의 움직임과 같은 시각적으로 뚜렷한 조음 동작이 포함된 문장을 확보하기 위한 것으로, 한국어의 양순 자음(ㅂ, ㅃ, ㅍ, ㅁ) 및 원순 모음(/ㅗ/, /ㅜ/)을 우선적으로 고려하였다. 모음의 경우, 혀의 고저에 따라 턱의 움직임이 달라지며, 이중모음 또한 입술과 턱의 복합적인 움직임에 영향을 미치기 때문에, 다양한 모음이 고르게 분포된 문장을 포함하는 것이 중요하다.

모음 분포의 균형을 위해 신지영의 한국어 발화 모음 빈도 자료[23]를 참고하였으며, 이를 바탕으로 1,000문장 단위로 모음의 출현 비율을 설정하여 문장을 선정하였다. 이 과정에서는 사용 빈도가 높은 모음이 더 자주 등장하도록 하되, 양순 자음과의 결합 여부를 조건으로 설정하여 조음 동작의 다양성을 확보하였다. 특히 양순 자음은 치경음이나 연구개음에 비해 출현 빈도가 낮기 때문에, 해당 자음이 배제되지 않도록 별도의 기준을 마련하였다.

기준 ②는 위에서 선정된 음운들이 문장의 특정 위치에만 집중되지 않도록 하기 위한 것으로, 문두와 문중에서 해당 음운이 고르게 분포되도록 하였다. 이를 통해 생성형 AI 학습 시 다양한 조음 위치의 데이터를 확보할 수 있도록 하였다.

기준 ③과 ④는 문장의 자연스러움과 표현의 다양성을 보장하기 위한 것으로, 실제 한국어 화자가 사용하는 문장을 중심으로 구어 및 문어 코퍼스를 활용하였다. 구어체 문장은 실제 발화 환경을 반영할 수 있으며, 문어체 문장은 문형의 확장을 가능하게 한다. 이에 따라 종결 어미의 다양성, 문장의 복잡도 등을 고려하여 다양한 문장 유형이 포함되도록 구성하였다.

기준 ⑤는 데이터 활용 목적에 부합하는 발화 시간의 일관성을 확보하기 위한 것이다. 음절 수 기준으로는 문장당 2030음절 사이에서 선정하였으며, 1,000문장 단위로 평균 2224음절의 길이를 유지하도록 하였다. 이를 통해 녹음 및 후속 처리 과정의 효율성을 높일 수 있었다.

조건을 두루 반영하고 있는 문장을 생성하기 위해 1,000문장을 단위로 DB, 작업 목표를 달리하였다. 단위별 작업 현황을 정리하면 [표 1]과 같다.

[표 3] 단위별 대상 자료 및 목표 모음 등장 위치

단위	DB		작업 목표	선정 문장 수
1	구어	독백	한국어 음운이 모두 실현될 것	1,000개
2		대화	문중에 목표 음운들이 위치할 것	1,007개
3			문두에 목표 음운들이 위치할 것	1,007개
4	문어		문중에 목표 음운들이 위치할 것	1,000개
5			문두에 목표 음운들이 위치할 것	1,000개

문장을 추출하기 위해 이 연구에서는 구어 자료와 문어 자료를 모두 활용하였다. 구어는 음성을 매체로 청각 채널을 통해 전달되는 것이라면, 문어는 문자를 매체로 시각 채널을 통해 전달되는 것이다. 구어의 호흡이라는 생리적 한계를 가지며, 즉각적으로 이루어진다는 점에서 문어와 차이가 있다.[24] 이러한 차이로 인해 구어냐 문어냐에 따라 자주 관찰되는 어휘와 표현이 다르고, 발화 반복, 담화표지 등 문어에서 관찰되지 않는 언어 현상이 구어에서 관찰되기도 한다.

이 연구에서는 이러한 차이를 고려하여 구어와 문어 자료를 모두 활용하였고, 1000분장을 단위로 작업마다 작업 코퍼스를 다르게 사용하였다. 아래의 [표 2]는 대본에 활용한 자료의 유형과 사용 코퍼스를 정리한 것이다.

[표 4] 대본에 활용한 자료 유형 및 사용 코퍼스

단위	자료 유형		사용 코퍼스
1	구어	독백 자유 발화	2차 말글 비교 실험 DB (50명, 221분 분량, 본 연구팀 구

23) 신지영. "성인 자유 발화 자료 분석을 바탕으로 한 한국어의 음소 및 음절 관련 빈도." 언어청각장애연구(한국언어청각임상학회) (2008): 13-2.

24) 신지영. "구어 연구와 운율: 소리를 담은 의미· 통사론, 의미를 담은 음성학· 음운론 연구를 위한 제언." 한국어 의미학 44 (2014): 119-139.

			축)
2 3		대화	2인 대화 DB (276명, 약 2,760분, 본 연구팀 구축)
4 5	문어	신문 기사	〈21세기 세종계획 중 신문 DB〉

구어 자료는 독백 자유 발화와 대화 발화로 나누었다. 독백 자유 발화는 일정한 주제를 갖고 발표하듯이 이루어진 발화로 고려대학교 음성언어정보연구실에서 구축한 '2차 말글 비교 실험 DB'[25]를 사용하였다. '2차 말글 비교 실험 DB'는 구어와 문어를 직접적으로 비교하기 위한 목적으로 구축된 것으로, 동일인이 유사한 담화 상황에서 동일한 주제에 대해 수업 시간에 발표하듯이 말하고, 보고서를 작성하듯이 글을 쓴 자료를 수집하는 방식으로 구어와 문어 자료가 각각 수집되었다. 실험 참가자는 20-30대 화자 204명으로, 하나의 주제에 대해 5분 내외의 독백 발화를 수행하였다. 대본 구축에 직접적으로 활용한 것은 총 50인의 발화 자료, 약 221분 분량이다.

다음으로 대화 발화는 미리 정해진 주제나 대본 없이 2인 이상의 화자가 자유롭게 대화한 것으로 고려대학교 음성언어 정보연구실에서 구축한 '해-해 DB', '해-해요 DB', '해요-해요 DB'를 사용하였다. '해-해 DB'는 모두 해체를 사용하는 두 명의 화자가 20분간 특정한 주제 없이 자유롭게 발화한 것이다. 화자들은 서로 친밀한 동갑의 친구 관계로, 동성 화자와 한 차례, 이성인 화자와 한 차례 20분씩 대화하는 방식으로 구성되었다. 규모는 40개 대화 자료, 약 800분 분량이다. '해-해요 DB'는 상하 관계에 있는 성인 화자 2인의 20분 대화 자료로 한쪽은 반말을, 한쪽은 존댓말을 사용하여 대화한 것이다. 연소자 남성과 연장자 남성, 연소자 남성과 연장자 여성, 연소자 여성과 연장자 남성, 연소자 여성과 연장자 여성이 약 15쌍씩 유지되도록 자료를 구축하였다. 규모는 68개 약 1,360분 분량이다. '해요-해요 DB'는 여성이 초면의 또래 동성과 20분간 대화한 것, 남성이 초면의 또래 동성과 20분간 대화한 것, 초면의 또래 이성과 20분간 대화한 것 10개로 규모는 30개 대화 자료, 약 600분 분량이다.

문어 자료로는 21세기 세종계획 현대국어 기초말뭉치[26]를 활용하였다. 이는 1998년부터 2006년까지의 문학작품, 신문 기사, 교양 서적 등 각종 현대 국어 문어 자료이다. 이 데이터셋 구축 연구에서는 소설과 같이 대화적 상황을 고려하여 작성된 문어 자료를 제외하고자 신문 기사로만 자료를 제한하여 6개 언론사(경향신문, 동아일보, 조선일보, 중앙일보, 한겨레신문, 한국일보)의 기사 자료를 활용하였다.

최종적으로, 5단위에 걸쳐 5,014개의 문장을 선정하였다. 1단위는 독백 자유 발화 코퍼스에서 추출된 1,000개의 문장이다. 2단위와 3단위는 2인 대화 코퍼스를 대상으로 하였는데, 문중 위치 1,007개, 문두 위치 1,007개가 선정되었다. 4단위와 5단위는 세종 코퍼스 신문 기사 자료에서 선정된 문중 위치 1,000개, 문두 위치 1,000개 문장이다. 아래의 〈표 3〉는 단위별 작업 개요 및 선정된 문장의 일련번호를 정리한 것이다. 결과물은 별도의 엑셀 파일로 제출한다.

[표 5] 단위별 작업 개요 및 문장 일련번호

단위	개요	일련번호
1	독백 자유 발화	1~1000
2	대화(구어) 문중 위치	1001~2007
3	대화(구어) 문두 위치	2008~3014
4	신문 기사(문어) 문중 위치	3015~4014
5	신문 기사(문어) 문두 위치	4015~5014

3. 전체 구축 과정의 연구

본 연구에서는 한국어 발화 기반 3D 디지털 휴먼(가상인간) 모델 생성을 위한 데이터 구축 및 처리 과정을 설계, 수집, 정제, 가공, 검수, 학습 모델 구현, 법적 권리 확보의 일련의 절차에 따라 수행하였다. 본 절에서는 각 단계에서의 구체적인 실행 내용을 통합적으로 기술한다.

먼저, 데이터 설계 단계에서는 음성 기반 3D 발화 얼굴 제작과 관련된 기존 자료를 검토하였다. 이를 통해 한국어에 특화된 관련 데이터가 현저히 부족하다는 점을 확인하였으며, 자연스러운 입모양 구현을 위해 요구되는 데이터의 범위와 규모를 산정하

25) 신지영, and 김경화. "한국인 표준 음성 DB 구축 (II)." 말소리와 음성과학 9.2 (2017): 9-22.

26) 김흥규, 강범모, and 홍정하. "21 세기 세종계획 현대국어 기초말뭉치: 성과와 전망." 한국정보과학회 언어공학연구회 학술발표 논문집 (2007): 311-316.

였다. 또한, 데이터 수집 및 공개 시 발생할 수 있는 법적 리스크를 사전에 차단하기 위해 법무법인을 통한 법률 자문을 실시하고, 데이터 구축을 위한 주요 조건들을 정의하였다.

이후, 원시데이터 수집은 오모션[27] 페이셜 캡처 스튜디오에서 진행되었으며, 총 5,000개의 문장을 대상으로 10명의 발화 모델(아나운서 및 배우)을 활용하여 다시점 촬영을 수행하였다. 각 모델은 500문장씩 발화하였으며, 모든 문장은 평균 발화 시간 약 3.8초로 구성되어 언어학적 기준에 맞추어 작성되었다. 촬영에 앞서, 각 모델에게 데이터 수집 및 활용, 제3자 제공에 대한 설명을 제공한 후, 개인정보 수집 및 이용에 대한 동의를 포함한 서면 동의서를 확보하였다.

수집된 다시점 영상 데이터로부터는 3차원 합성을 위한 이미지 시퀀스를 추출하였고, 텍스처 생성용 이미지로도 변환하였다. 문장당 50개 시점에서 촬영된 영상 중 정면 카메라 영상을 기준으로 오디오 데이터를 추출하여, 음성-영상 정렬에 필요한 기초 데이터를 확보하였다.

다음 단계에서는 원천데이터 가공 작업이 이루어졌다. 가공 데이터는 (1) 3D 메쉬, (2) 입 부위의 3D 메쉬 버텍스, (3) 3D 얼굴 랜드마크로 구성되었다. 50개 시점의 이미지 기반으로 고해상도 3D 메쉬를 생성하였으며, 입 부위 버텍스는 사전에 정의된 index 기준에 따라 추출되었다. 전체 메쉬는 약 24,000개의 버텍스로 구성되었으며, 이를 리토폴로지(retopology) 과정을 통해 경량화하고, 동일한 위치가 동일한 의미를 갖도록 표준화하였다. 또한, 학습 및 리토폴로지를 위한 68개의 3D 얼굴 랜드마크를 라벨링 기법을 통해 생성하였다.

가공 데이터를 대상으로는 체계적인 검수 절차를 적용하였다. 데이터 구축의 각 단계마다 품질 검수 기준을 수립하고, 데이터 유형별(예: 랜드마크, 립 버텍스, 3D 메쉬)에 대해 육안 검수 도구와 코드 기반 자동 검수 도구를 함께 활용하였다. 검수는 단계별로 배정된 전문 인력이 수행하였으며, 특히 3D 메쉬의 경우 크라우드워커 기반 체크리스트를 활용한 전수 검사를 통해 정밀하게 검토되었다.

학습모델 구현은 기존의 대표적인 음성-얼굴 연동 모델인 CodeTalker와 FaceFormer를 비교한 후, 다양한 얼굴 움직임을 더 효과적으로 모델링할 수 있는 CodeTalker를 기반으로 진행되었다. CodeTalker는 얼굴의 움직임 공간(motion space)을 codebook으로 표현하는 방식을 제안하며, 이는 연속적인 얼굴 동작을 유한한 이산적 표현으로 양자화한 것이다. 구체적으로는 트랜스포머 기반의 VQ-VAE(Vector Quantized Variational AutoEncoder)를 학습하여 codebook을 생성하고, 해당 codebook을 추론에만 사용하였다. 이 방식은 codebook을 활용하여 얼굴의 움직임 강도나 정적·동적 특성을 조절할 수 있다는 점에서 높은 유연성과 사용자 정의 가능성을 제공한다.

마지막으로, 법적 권리 확보를 위한 조치도 병행되었다. 본 연구에서 수집한 데이터에는 발화자의 얼굴 영상뿐만 아니라, 연령, 성별 등 개인정보 및 민감정보가 포함되므로, 수집 및 공개와 관련된 저작권, 초상권, 개인정보보호법 등 다양한 법적 요소에 대해 법률 자문을 거쳤다. 특히, 비식별화가 어려운 얼굴 영상 데이터를 다루는 만큼, 수집 및 이용, 제3자 제공, 초상권 활용 등에 대한 명확한 동의가 필요하다는 자문 결과에 따라, 이를 모두 포함한 통합 동의서를 제작하여 모델에게 서면 동의를 받았다.

4. 데이터 세트의 구축

4-1. 다시점 비디오의 촬영

본 연구에서는 음성 기반 3D 발화 얼굴 생성을 위한 데이터 구축을 목적으로, 관련 자료 조사를 통해 한국어에 특화된 3D 발화 얼굴 데이터가 거의 존재하지 않는다는 문제를 확인하였다. 이에 따라, 자연스러운 한국어 입모양 구현을 위해 필요한 데이터의 규모와 세부 요건을 설정하고, 총 5,000문장의 데이터를 수집하기로 하였다. 수집 대상은 10명의 발화 모델(아나운서 및 배우)이며, 각 모델은 500문장씩 다시점 카메라 환경에서 촬영을 진행하였다. 모든 문장은 평균 3.8초의 길이를 갖도록 구성되었으며, 음운적 다양성과 언어학적 구조를 고려하여 설계되었다. 또한, 데이터 수집 및 공개 과정에서의 법적 문제를 사전에 방지하기 위해 법무법인의 법률 자문을 거쳐 개인정보 보호, 초상권, 저작권 등에 대한 검토를 수행하였고, 발화 모델로부터 관련 동의서를 수집하였다.

데이터 수집은 총체적으로 다음과 같은 절차에

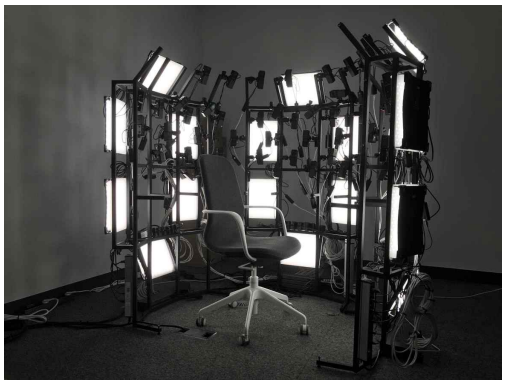
27) http://omotion.co.kr/

따라 수행되었다. 먼저, 수집 환경을 구축하는 단계에서는 고품질의 원시데이터 확보를 위한 조명, 촬영 장비, 스튜디오 설비 등이 마련되었다. 조명 시스템은 앰비언트 조명을 기반으로 구성하여 다시점 촬영 시 음영이 발생하지 않도록 하였으며, 낮은 ISO 설정에서도 충분한 조도를 확보할 수 있도록 하였다. 카메라 리그는 다양한 각도에서 얼굴을 정확히 포착할 수 있도록 고정 설치되었으며, 네트워크 및 전원 케이블의 정렬을 통해 안정적인 유지보수가 가능하도록 구성하였다. 촬영 스튜디오는 외부 광원의 영향을 최소화하고, 음향 녹음 품질 저하를 방지하기 위해 격리된 구조로 설치하였다. 또한, 전문 촬영 가이드를 통해 발화자의 표정이 명확하게 드러나도록 지도하였고, 피부톤 보정 및 3D 스캔 오류 방지를 위한 메이크업 등 촬영 준비 과정을 사전에 진행하였다.

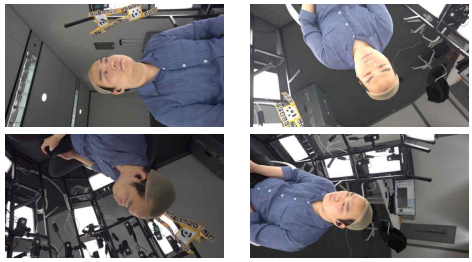
데이터 수집은 각 모델에게 500개의 문장으로 구성된 대본을 제공하고, 목표 음운을 명확하게 발음할 수 있도록 발화 시 유의사항을 사전 전달하는 방식으로 이루어졌다. 발화자는 사전 안내된 내용을 바탕으로 각 문장을 자연스럽게 발화하였으며, 이를 통해 총 5,000개의 한국어 발화 데이터를 수집하였다.

촬영은 다시점 영상 방식으로 진행되었으며, 전용 소프트웨어를 활용하여 50대의 카메라를 동기화하고 제어함으로써 발화자의 얼굴을 다양한 시점에서 고화질로 촬영하였다. 각 카메라는 1920×1080 해상도, 60P50M, 30fps의 스펙으로 설정되었으며, 얼굴의 미세한 움직임까지 정밀하게 포착할 수 있도록 조명과 위치를 정교하게 조정하였다. 카메라의 동기화를 통해 영상 간 시간 일치를 유지함으로써 데이터의 정밀도를 확보하였다.

촬영된 영상은 모델 및 스크립트별로 체계적으로 분류 및 저장되었으며, 이후 데이터 분석 및 3D 모델 가공을 위한 기초 자료로 활용될 수 있도록 안정적인 보관 체계를 갖추었다. 최종적으로, 10명의 모델 × 500문장 × 50시점이라는 구조로 구성된 대규모 원시 데이터를 구축하였다.



[그림 1] 본 연구에 사용된 다시점 촬영 시스템



[그림 2] 다시점 촬영 결과의 예시

4-2. 영상 및 오디오 데이터의 처리

본 절에서는 구축된 다시점 원시 영상으로부터 이미지 프레임 및 오디오 데이터를 추출하고 정제하는 과정과 품질 검수 및 저장 절차를 상세히 기술한다. 해당 데이터는 이후 3D 메쉬 합성 및 생성형 AI 학습의 기반 자료로 활용되며, 일부는 외부 검증을 위해 제출되었다.

먼저, 원시 영상으로부터 이미지 프레임을 추출하기 위해 자체 개발한 전용 소프트웨어를 사용하였다. 이 소프트웨어는 영상 데이터로부터 모든 프레임을 고속으로 분리하여 저장할 수 있도록 설계되었으며, 이미지 추출 성능을 향상시키기 위해 OpenMP[28] 기반 멀티스레딩 기술을 적용하였다. 이로써 대용량의 고해상도 영상 데이터로부터 빠르고 안정적으로 이미지 프레임을 확보할 수 있었다.

이미지 추출 후에는 내부 품질 검수 인력을 활용하여 이미지 프레임에 대한 시각적 검수를 진행하였다. 검수 항목은 해상도, 촬영 각도, 조명 조건 등을 포함하며, 이 과정을 통해 확보된 이미지 데이터의 품질을 체계적으로 보장하였다. 검수가 완료된 이미

28) https://www.openmp.org/

지 데이터는 문장당 50장의 프레임 단위로 묶어 저장되었으며, 각 이미지의 파일명은 촬영 카메라의 명칭과 프레임 번호를 조합하여 구성되었다. 또한, 저장된 이미지에는 촬영 일자, 촬영자 등 주요 메타데이터를 함께 기록하여 후속 데이터 분석과 관리의 효율성을 높였다.

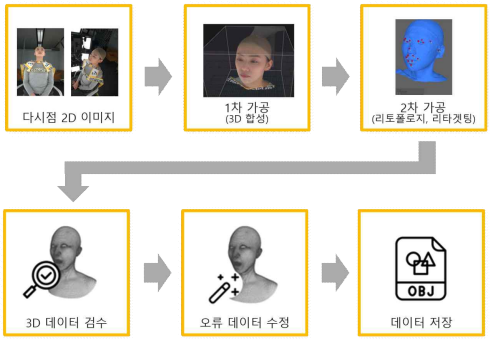
결과적으로, 약 250,000개의 다시점 영상 클립(총 5,000문장 × 50시점)으로부터 약 28,500,000장의 2D 이미지 프레임을 추출하였다(평균 3.8초, 30fps 기준). 이 데이터는 주로 3D 메쉬 합성을 위한 입력 자료로 사용되며, 생성형 AI 모델 학습에는 활용되지 않는 중간산물로 분류된다. 이에 따라, AI Hub에는 제출하지 않되, 품질 검증을 위해 정면 방향에서 촬영된 약 570,000장의 이미지 데이터는 한국정보통신기술협회(TTA)[29]에 제출하였다.

한편, 오디오 데이터는 다시점 영상 중 정면 카메라로 촬영된 영상으로부터 추출되었다. 원본 영상은 MP4 포맷이며, 음성은 WAV 포맷으로 변환하여 저장하였다. 추출된 오디오 데이터는 전담 검수 인력을 통해 품질 확인 및 영상과의 동기 여부를 검수한 후, 각 발화 모델과 문장 단위로 체계적으로 분류 및 저장되었다. 최종적으로, 10명의 발화 모델 × 500문장 구성에 따라 총 5,000개의 고품질 오디오 데이터를 확보하였다.

4-3. 3차원 얼굴 데이터의 제작

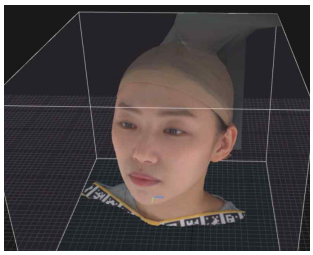
3D 발화 얼굴 데이터 생성을 위해, 다시점 이미지로부터 얼굴의 주요 특징점을 추출하고 이를 바탕으로 point cloud를 생성하였다. 이 과정에서는 어깨부터 얼굴 상부까지의 영역을 복원 대상으로 설정하였으며, 다시점 이미지 정합(multi-view image alignment) 기술을 적용하여 정밀한 3차원 데이터 생성을 수행하였다. 생성된 3D 데이터에는 학습 목적의 활용도를 높이기 위해 얼굴의 주요 랜드마크를 라벨링하여 포함시켰다. 이후, 약 24,049개의 버텍스를 가진 고해상도 템플릿을 기준으로, 생성된 3D 데이터와 각 랜드마크를 고정점(anchor point)으로 활용하여 매핑(wrapping)하고, 리토폴로지(retopology)를 수행하였다. 이 과정은 데이터의 정합성뿐 아니라 경량화와 학습 효율성을 동시에 확보하기 위한 핵심 절차이다.

최종적으로는 10명의 모델이 각각 500개의 문장을 발화하고, 3.8초 길이의 데이터를 30fps로 촬영한 결과, 570,000개 이상의 3D 메쉬 데이터를 구축하였다. 이 데이터는 어깨~얼굴 영역의 고정밀 3D 복원을 포함하고 있으며, 리토폴로지 과정은 랜드마크 기반의 정합을 통해 이루어졌다.



[그림 3] 데이터 가공 절차

이와 함께, 3D 데이터의 구조적 정보 외에 메타데이터 가공도 병행되었다. 메타데이터는 원시데이터 수집 단계에서 확보한 기본 정보(촬영일시, 발화 모델 정보 등)와 3D 가공 과정에서 생성된 라벨링 정보를 통합하여 구성하였다. 생성된 메타데이터는 각 3D 메쉬 데이터와 1:1로 매칭되었으며, 따라서 총 570,000개 이상의 세트가 제작되었다. 해당 메타데이터는 향후 데이터 검색, 관리, 분석을 용이하게 하기 위한 필수 기반 자료로 활용된다. 메타데이터의 구체적 구성 항목은 표 ①에 정리하였다.

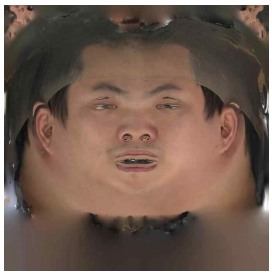


[그림 4] 3D 얼굴의 합성 결과

또한, 3D 메쉬와 함께 텍스처 데이터도 고해상도로 가공되었다. 초기 생성된 텍스처는 비정렬 상태였으며, 이를 정렬된 레퍼런스 텍스처 UV 맵과 매칭시켜 눈, 코, 입, 눈썹 등 얼굴의 주요 시각적 특

29) https://www.tta.or.kr/

징이 명확히 드러나도록 재구성하였다. 텍스처 해상도는 2048×2048 수준으로 설정하여 발화자의 얼굴 특성을 최대한 반영하였으며, 모델 1인당 1개의 텍스처를 생성하여 총 10개의 고해상도 텍스처가 구축되었다.



[그림 5] 텍스쳐 데이터 예시

4-4. 데이터 학습 및 테스트

본 연구에서는 3D 발화 얼굴 생성을 위한 학습 모델로 CodeTalker 구조를 기반으로 한 2단계 학습 방식을 설계하였다. CodeTalker는 얼굴의 움직임을 정량적으로 표현할 수 있는 codebook을 활용하여, 음성 입력으로부터 자연스러운 얼굴 표정을 생성하는 데 최적화된 모델이다. 학습은 크게 두 단계(Stage 1, Stage 2)로 구분하여 수행되었다.

Stage 1은 모델이 표현 가능한 얼굴 움직임의 전체 범위를 학습하는 단계이다. 이 단계에서는 음성 입력을 사용하지 않고, 다양한 얼굴 표현을 기반으로 움직임의 공간(motion space)을 정의하고 이를 유한한 수의 이산적 표현으로 양자화(quantization)하는 과정을 수행하였다. 구체적으로는 VQ-VAE(Vector Quantized Variational AutoEncoder) 기반의 구조를 활용하여 codebook을 학습하였으며, 이 codebook은 얼굴의 다양한 움직임을 대표하는 이산적인 코드들로 구성된다. 결과적으로, Stage 1에서는 입력된 얼굴 시퀀스에서 추출된 고차원 움직임을 정제된 정수 인덱스로 변환하는 기초 모델이 형성되었다.

Stage 2는 음성 입력을 기반으로, 해당 발화에 적합한 얼굴 움직임 코드(codebook index)를 예측하는 단계이다. 이 단계에서는 transformer 기반 구조를 도입하여 음성 신호와 얼굴 움직임 간의 복잡한 관계를 모델링하였다. 음성의 시간적 패턴과 주파수 특징이 어떻게 얼굴의 표정 변화에 반영되는지를 학습함으로써, 보다 자연스럽고 일관된 입모양과 표정 애니메이션을 생성할 수 있도록 하였다. 즉, Stage 2는 주어진 음성에 대하여 codebook 내 어떤 얼굴 움직임 인덱스를 순차적으로 출력할지를 결정하는 예측 모델을 학습하는 과정이다.

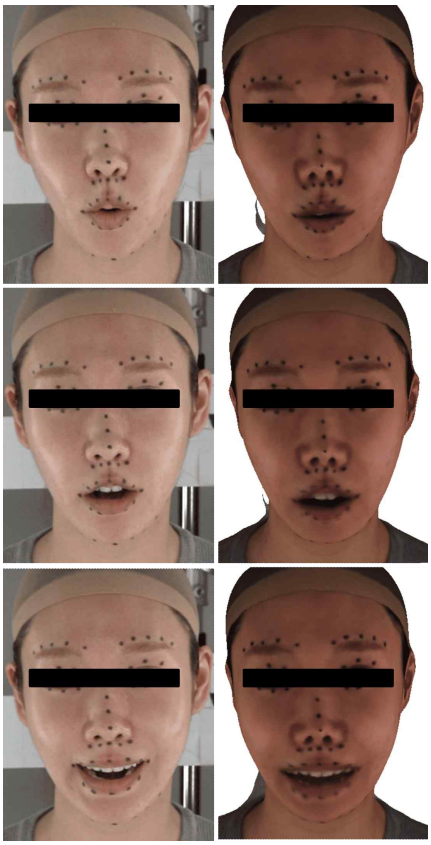
이러한 이단계 학습 구조는 음성과 얼굴 움직임의 상관관계를 효과적으로 학습할 수 있을 뿐만 아니라, 실제 생성 시에도 표현력 있는 디지털 휴먼의 시각적 구현을 가능하게 하며, 고품질의 발화 영상 합성에 기여한다.

5. 데이터 구축 결과

본 연구에서는 음성 기반 3D 발화 얼굴 생성에 필요한 다양한 형태의 데이터셋을 구축하였다. 데이터는 원천데이터와 라벨링 데이터의 두 범주로 구분되며, 각각의 항목은 3D 메쉬 생성과 학습 모델 구축에 필수적인 정보를 포함한다.

원천데이터는 다시점 2D 이미지 데이터와 오디오 데이터로 구성된다. 먼저, 다시점 2D 이미지 데이터는 3D 메쉬 생성을 위해 다시점 영상에서 프레임 단위로 추출한 이미지로, 총 570,000개 이상의 PNG 형식의 이미지가 구축되었다. 이 중 정면방향의 이미지는 3D 메쉬 품질 검증을 위해 활용되며, 3D 메쉬 가공 과정에서 파생된 중간 산출물이므로 AI Hub에는 비공개로 처리되었다. 오디오 데이터는 각 발화 영상으로부터 추출한 음성으로, 총 5,000개의 WAV 포맷 파일이 저장되었다. 해당 오디오 파일은 영상과의 싱크를 맞추어 모델별 및 문장별로 정리되었다.

라벨링 데이터에는 3D 메쉬 데이터, 텍스처 데이터, 그리고 메타데이터가 포함된다. 3D 메쉬 데이터는 다시점 2D 이미지를 바탕으로 모델링된 3D 형상으로, 각 메쉬는 24,049개의 버텍스(Vertices)로 구성되어 있으며, 총 570,000개 이상의 OBJ 포맷 파일이 생성되었다. 텍스처 데이터는 모델별 텍스처 파일로서, 2048×2048 해상도의 PNG 이미지로 구성되며, 총 10개의 텍스처가 구축되었다.



[그림 6] 구축 데이터 예시

마지막으로, 메타데이터는 모델 정보, 촬영 환경 정보와 더불어 리토폴로지 및 학습에 필요한 라벨 정보로 구성된다. 주요 항목에는 얼굴 랜드마크 정보(68 포인트)와 입술 영역 버텍스의 인덱스 및 좌표 정보(총 4,410개의 버텍스)가 포함되어 있으며, 총 570,000개 이상의 JSON 파일이 생성되었다. 이 메타데이터는 각 3D 메쉬 파일과 1:1로 매칭되어 활용 가능하도록 설계되었다.

[표 6] 음성 기반 3D 발화 얼굴 데이터 구축 세부 내역

구분	데이터명	설명	포맷	구 축 수 량
원천 데이터	다시점 2D 이미지 데이터 - 정면 이미지	3D 메쉬 데이터 생성을 위해 다시점 영상에서 추출한 이미지 프레임(다시점 이미지 중 3D 메쉬 품질 검증을 위해 정면방향의 이미지를 활용함, 3D 메쉬 가공 과정에서 산출되는 부산물로 AIhub에 공개하지 않음)	PNG	578,242개
	오디오 데이터	영상에서 추출한 오디오	WAV	5,000개
라벨링 데이터	3D 메쉬 데이터	다시점 2D 이미지를 3D 모델로 모델링한 데이터 24,000 vertices로 구성	OBJ	570,000개 이상
	텍스쳐 데이터	모델별 텍스쳐 파일 해상도: 2048x2048	PNG	10개
	메타 데이터	모델 및 촬영 메타 정보, 리토폴로지 시 사용한 얼굴 랜드마크 정보(68 points), Lip vertex의 인덱스 및 좌표 정보(4,410 vertices)	JSON	570,000개 이상 (3D메쉬 데 이 터와 1:1 매칭)

6. 결론

본 연구는 자연스러운 한국어 발화를 구현할 수 있는 3D 디지털 휴먼 제작을 목표로, 음성 기반 3D 발화 얼굴 데이터를 구축하고 이를 활용한 생성형 인공지능 학습 모델을 설계·개발하였다. 조음 특성과 음운 분포를 고려한 5,000문장 대본을 바탕으로 다시점 촬영을 통해 570,000개 이상의 PNG 이미지와 5,000개의 WAV 파일을 확보하였으며, 총 570,000개 이상의 고해상도 OBJ 파일과 JSON 메타데이터, 24,049개의 버텍스로 구성된 정밀 3D 메쉬를 구축하였다. 이를 기반으로 트랜스포머 구조의 CodeTalker 모델을 학습하여, 음성과 표정 간의 정합성이 높은 고품질 3D 얼굴 애니메이션을 구현하였다.

이와 같은 데이터 구축과 모델 개발 과정은 음성과 3D 애니메이션의 정밀한 연동을 실현함으로써, 한국어 특화 디지털 휴먼 제작의 기술적 기반을 공고히 하였으며, 학문적으로는 음성 기반 3D 표현 연구의 새로운 사례를 제시하였다는 데 의의가 있다. 특히 기존에 존재하지 않았던 한국어 전용 3D 발화 데이터셋을 체계적으로 설계·수집·가공한 점은 향후 관련 연구와 기술 발전의 출발점으로 기능할 수 있다.

연구 성과는 실용적 응용 가능성 또한 높다. 생성형 AI 기반 디지털 휴먼 발화 생성 서비스에 바로 활용 가능하며, 방송 콘텐츠 제작, 메타버스 기반 상호작용형 아바타, 공공분야 AI 챗봇 등 다양한 산업

영역에서 응용될 수 있다. 특히 고품질 입모양과 표정을 구현함으로써 콘텐츠 몰입도를 높이고, 제작 효율성을 크게 향상시킬 수 있다.

향후에는 데이터와 모델의 유지관리 체계를 지속 운영하고, 소스코드와 사용 매뉴얼을 공개함으로써 국내 연구자 및 개발자들과의 협업과 기술 확산을 도모할 계획이다. 본 연구는 음성 기반 3D 디지털 휴먼 기술의 국내 자립화에 기여하고, 인공지능과 메타버스가 융합되는 차세대 콘텐츠 산업에서 핵심적인 역할을 수행할 수 있을 것으로 기대된다.

마지막으로, 본 연구는 한국어 발화와 3D 얼굴 애니메이션을 결합한 새로운 기술을 제시하며, 가상인간 및 디지털 휴먼 기술의 발전에 중요한 기여를 할 것이다. 특히, 한국어에 최적화된 3D 얼굴 데이터셋을 구축하고, 이를 기반으로 자연스러운 얼굴 애니메이션을 생성하는 방법을 제시한다. 이 연구는 가상인간의 상호작용 능력을 한층 더 향상시키고, 다양한 분야에서의 활용 가능성을 넓히는 데 중요한 역할을 할 것이다. 또한, 한국어와 같은 비영어권 언어에 대한 연구는 글로벌 기술 개발에 중요한 기여를 할 수 있으며, 향후 다른 언어로의 확장 가능성도 열어줄 것이다.

참고문헌

1. 오문석, 한규훈, and 서영호. "메타버스를 위한 디지털 휴먼과 메타휴먼의 제작기법 분석 연구." 한국디자인리서치학회 6.3, 2021
2. 강석찬, and 김동주. "딥러닝을 활용한 한국어 스피치 애니메이션 생성에 관한 고찰." 정보처리학회논문지. 소프트웨어 및 데이터 공학 12.10, 2023
3. 김현주, et al. "고해상도 3D 데이터 생성 기술 분석 및 연구 동향." [ETRI] 전자통신동향분석 37.3, 2022
4. Ming, Xin, et al. "High-Quality Mesh Blendshape Generation from Face Videos via Neural Inverse Rendering." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
5. Cudeiro, Daniel, et al. "Capture, learning, and synthesis of 3D speaking styles." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
6. Wuu, Cheng-hsin, et al. "Multiface: A dataset for neural face rendering." arXiv preprint arXiv:2207.11243, 2022
7. Fanelli, Gabriele, et al. "A 3-d audio-visual corpus of affective communication." IEEE Transactions on Multimedia 12.6, 2010
8. Zhang, Jie, and Robert B. Fisher. "3d visual passcode: Speech-driven 3d facial dynamics for behaviometrics." Signal processing 160, 2019
9. Sung-Bin, Kim, et al. "MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset." arXiv preprint arXiv:2406.14272, 2024
10. Sung-Bin, Kim, et al. "Laughtalk: Expressive 3d talking head generation with laughter." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
11. Liu, Chang, et al. "EmoFace: Audio-driven emotional 3D face animation." 2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR). IEEE, 2024.
12. Xu, Zhihao, et al. "KMTalk: Speech-Driven 3D Facial Animation with Key Motion Embedding." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
13. Xing, Jinbo, et al. "Codetalker: Speech-driven 3d facial animation with discrete motion prior." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
14. Sha, Xuanmeng, et al. "3DFacePolicy: Speech-Driven 3D Facial Animation with

Diffusion Policy." arXiv preprint arXiv:2409.10848, 2024

15. Shen, Kang, et al. "DEITalk: Speech-Driven 3D Facial Animation with Dynamic Emotional Intensity Modeling." Proceedings of the 32nd ACM International Conference on Multimedia. 2024.
16. Peng, Ziqiao, et al. "Emotalk: Speech-driven emotional disentanglement for 3d face animation." Proceedings of the IEEE/CVF international conference on computer vision. 2023
17. Sun, Zhiyao, et al. "Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models." ACM Transactions on Graphics (TOG) 43.4, 2024
18. Lin, Yihong, et al. "Gliditalker: Speech-driven 3d facial animation with graph latent diffusion transformer." arXiv preprint arXiv:2408.01826, 2024
19. Peng, Ziqiao, et al. "Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces." Proceedings of the 31st ACM International Conference on Multimedia. 2023
20. 장민정, 정선진, and 노준용. "한국어 동시조음 모델에 기반한 스피치 애니메이션 생성." 한국컴퓨터그래픽스학회논문지 26.3, 2020
21. 강석찬, and 김동주. "딥러닝을 활용한 한국어 스피치 애니메이션 생성에 관한 고찰." 정보처리학회논문지. 소프트웨어 및 데이터 공학 12.10, 2023
22. 신지영. "성인 자유 발화 자료 분석을 바탕으로 한 한국어의 음소 및 음절 관련 빈도." 언어청각장애연구](한국언어청각임상학회), 2008
23. 신지영. "구어 연구와 운율: 소리를 담은 의미· 통사론, 의미를 담은 음성학· 음운론 연구를 위한 제언." 한국어 의미학 44, 2014
24. 신지영, and 김경화. "한국인 표준 음성 DB 구축 (II)." 말소리와 음성과학 9.2, 2017
25. 김흥규, 강범모, and 홍정하. "21 세기 세종계획 현대국어 기초말뭉치: 성과와 전망." 한국정보과학회 언어공학연구회 학술발표 논문집, 2007
26. omotion.co.kr
27. pixune.com/blog
28. www.openmp.org
29. www.tta.or.kr