한국컴퓨터그래픽스학회
Korea Computer Graphics Society Vol. 31, No. 3, P. 87~98

메쉬 구조 비종속적 음성 기반 3차원 발화 애니메이션 생성

서광균[1○] 차시헌[2○] 나현호[2] 이인엽[2] 노준용[2*]

[1]Flawless AI [2]한국과학기술원 비주얼미디어연구실

edward.seo@flawlessai.com {chacorp, nahyeonho1023, leeinyup123, junyongnoh}@kaist.ac.kr

Mesh-Agnostic Audio-Driven 3D Facial Animation Generation

Kwanggyoon Seo[1○] Sihun Cha[2○] HyeonHo Nah[2] Inyup Lee[2] Junyong Noh[2*]

[1]Flawless AI [2]KAIST, Visual Media Lab

요 약

본 연구에서는 임의의 메쉬 구조를 가진 3차원 얼굴 모델에 대응 가능한 음성 기반 얼굴 발화 애니메이션 생성 방법을 제안한다. 기존의 신경망 기반 연구는 템플릿 얼굴 모델의 메쉬 구조를 바탕으로 음성 입력과 애니메이션 간의 매핑을 학습하기 때문에 새로운 메쉬 구조를 가진 얼굴 모델에 적용하기 어렵다. 본 방법은 새로운 신경망 모델 Wav2Rig를 통해 메쉬 구조에 비종속적인 변형 전달 (mesh-agnoistic deformation transfer) 이 가능한 사전학습 신경망 모델의 잠재공간과 음성 입력 간의 매핑을 학습하여 새로운 형태와 구조의 3차원 얼굴 모델에 대한 발화 애니메이션 생성을 가능케 한다. Wav2Rig는 음성입력으로부터 해석 가능한 잠재코드 시퀀스를 생성하고, 사전학습 신경망 모델은 각 잠재코드를 기반으로 대상 얼굴 모델을 변형하여 발화 애니메이션을 생성한다. 제안 기법은 학습된 메쉬 구조에 대해 높은 정확도를 달성함은 물론, 기존 방법으로는 불가능했던 다양한 메쉬 구조의 얼굴 모델에도 안정적으로 발화 애니메이션을 생성할 수 있음을 실험을 통해 검증하였다. 또한, Wav2Rig에서 생성되는 표현 잠재코드는 직관적인 애니메이션 조작을 가능하게 하며, wav2vec 2.0에서 추출된 음성 신호 특징 중 중간 계층의 특징이 3차원 얼굴 발화 애니메이션 성능 향상에 효과적임을 추가 분석을 통해 확인하였다.

Abstract

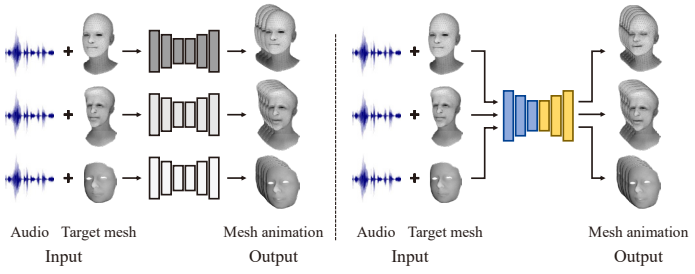
We present an end-to-end approach to animating a 3D face mesh with arbitrary shape and triangulation from a given speech audio. Previous approaches employ a neural network that can only drive a 3D face mesh in learned mesh structure, which limits the generalizability. To address this, our method leverages a pretrained mesh-agnostic deformation transfer model to enable the animation of a 3D face mesh with arbitrary shape and triangulation that are unseen during the training. Specifically, we design a mapping network Wav2Rig that produces a sequence of interpretable expression codes from the input audio. The pre-trained deformation network then deforms the target mesh according to the sequence of produced expression codes, resulting in a 3D facial animation. A set of experiments verifies that the proposed method achieves state-of-the-art results on the trained mesh topology and is capable of driving unseen 3D face meshes in different mesh topologies, which is impossible by previous methods. In addition, the interpretable expression codes from the Wav2Rig model enable easy manipulation of the generated facial animation. Lastly, we further delved into the analysis of audio features extracted from wav2vec 2.0 and found that the features from the middle layers of wav2vec 2.0 enhance the performance of audio-driven 3D facial animation.

○ Equal contribution, *Corresponding author

키워드: 음성 발화 애니메이션, 얼굴 애니메이션
Keywords: Speech Animation, Facial Animation

*corresponding author: Junyong Noh / KAIST, Visual Media Lab (junyongnoh@kaist.ac.kr)

Received : 2025.06.13./ Review completed : 1st 2025.06.30. 2nd 2025.07.10./ Accepted : 2025.07.14.
DOI : 10.15701/kcgs.2025.31.3.87
ISSN : 1975-7883(Print)/2383-529X(Online)



(a) Previous method – mesh specific (b) Our method – mesh agnostic

Figure 1: **Visual comparison of the methods.** The proposed method (b) can animate arbitrary 3D face meshes from audio input, without being constrained to a specific mesh structure. In contrast, previous methods [4, 5] (a) are limited by the mesh structure used during the training and require separate networks for each target mesh with a new mesh structure.

1 서론

3차원 얼굴 애니메이션은 TV 드라마의 음성-영상 더빙, 화상 회의용 디지털 아바타, 소셜 미디어, 라이브 스트리밍 등 다양한 분야에서 활용되어 왔다. 특히, 게임 및 영화 산업에서 3차원 콘텐츠의 수요가 높아짐에 따라, 몰입감 있는 스토리텔링을 위한 3차원 얼굴 발화 애니메이션의 중요성이 더욱 부각되고 있다. 한편, 기존의 3D 얼굴 애니메이션 제작 과정은 많은 시간과 숙련된 전문가의 작업을 필요로 하기 때문에 이러한 수요에 효과적으로 대응하기 어렵다. 이로 인해 작업 효율을 높이거나 자동화할 수 있는 솔루션에 대한 관심이 증가하고 있으며, 관련 연구 또한 활발히 이루어지고 있다.

최근에는 음성 신호를 활용하여 신경망을 통해 발화 애니메이션을 자동으로 생성하려는 연구가 활발히 진행되고 있다 [1, 2, 3, 4, 5, 6]. 해당 연구들은 공통적으로 입력된 음성 신호를 대상 얼굴 모델에서의 정점의 움직임으로 변환하도록 네트워크를 학습시킨다. 음성 신호를 3차원 얼굴 발화 애니메이션으로 직접 변환할 수 있다는 점에서, 과거에 숙련된 전문가의 많은 작업을 필요로 했던 제작 과정에 비해 전면 자동화가 가능하여 작업 효율을 크게 개선할 수 있다는 장점이 있다. 다만, 이러한 방법은 학습에 사용된 것과 동일한 메쉬 구조를 가진 얼굴 모델에만 적용 가능하다는 한계가 있다 (Fig. 1(a)). 다른 메쉬 구조의 얼굴 모델에 적용하기 위해서는 별도의 리타게팅 과정 또는 해당 얼굴 모델을 활용한 신경망의 재학습을 필요로 한다. 따라서 임의의 메쉬 구조를 가진 3차원 얼굴 모델에 대해 음성 기반 발화 애니메이션을 생성하는 것은 여전히 해결해야 할 과제로 남아 있다.

이러한 문제를 해결하기 위해, 본 연구에선 임의의 메쉬 구조를 가진 3차원 얼굴 모델에 대응 가능한 음성 입력 기반의 3차원 얼굴 애니메이션 생성 방법을 제안한다. 본 방법은 다양한 메쉬 구조를 가진 3차원 얼굴 모델에도 적용 가능한 메쉬 비종속적 변형 네트워크와 음성 인코더로 구성되며, 변형 네트워크로는 사전학습한 Neural Face Rigging (NFR) [7] 모델을 사용하고, 음성 인코더는 본 연구에서 새로이 제안하는 *Wav2Rig*을 활용한다. *Wav2Rig*는 사전학습한 음성 신호 분석 모델 wav2vec 2.0 [8]의 feature를 NFR의 표정 잠재코드 (expression latent code)로 변환하며, 해당 잠재코드는 Facial Action Coding System (FACS) [9]에 기반하여 정의된다. 생성된 표정 코드와 애니메이션 대상이 되는 3차원 얼굴 메쉬를 디코더에 입력하면, 메쉬의 변형 필드가 계산되어 음성 신호에 맞는 얼굴 애니메이션이 생성된다. 다양한 실험을 통해 제안하는 방법이 ICT [10], Multiface [11], VOCASET [12], BIWI [13] 등 서로 다른 메쉬 구조를 가진 데이터셋에서 적절한 음성 기반 얼굴 애니메이션을 생성할 수 있음을 실험을 통해 확인하였다.

추가적으로, 본 연구에서는 사전학습한 wav2vec 2.0 [8]의 계층 별 피처(feature)들을 분석한 후 *Wav2Rig* 학습을 위한 최적의 피처를 제시한다. 기존 연구들 [4, 5]은 사전학습한 wav2vec 2.0의 마지막 층에서 추출한 피처를 3차원 얼굴 애니메이션 생성을 위한 디코더의 입력으로 활용한다. 본 연구의 실험 결과에 따르면, wav2vec 2.0의 마지막 층이나 로짓 층에서 추출한 피처보다, 중간 계층(특히 5 ∼ 9번째 층)에서 추출한 피처를 사용하는 것이 더 높은 품질의 얼굴 애니메이션을 생성하는 데 효과적임을 확인하였다.

본 연구의 주요 기여 항목은 다음과 같이 정리할 수 있다:

- 특정한 얼굴 메쉬에 제한되지 않고, 입력 음성 신호만으로 3차원 얼굴에 대한 애니메이션을 생성할 수 있는 방법을 제안하였다.
- 음성 신호 기반 얼굴 애니메이션 생성을 위한 wav2vec 2.0 내부 피처 계층들에 대한 분석을 통해, 가장 효과적인 피처를 식별하였다.
- 음성 신호를 해석 가능한 표정 코드로 변환하는 *Wav2Rig*을 제안하였으며, 이를 통해 생성된 얼굴 애니메이션은 감정 조절이나 시각적 더빙과 같은 다양한 후작업(post-processing)에 활용 가능하다.

2 관련 연구

2.1 음성 기반 3차원 얼굴 발화 애니메이션

음성 신호 기반 3차원 얼굴 발화 애니메이션 생성은 Brand [14]의 초기 연구를 시작으로 오랜 기간 어려운 도전 과제로 남아 있었다. 해당 과제의 주요 목표는 음성 신호에 적절한 입모양을 지닌 발화 애니메이션을 생성하는 것이다. 심화학습 기반 신경망의 등장 이후, Audio2Face [1]는 음성 신호를 학습한 얼굴 메쉬의 정점 위치 값으로 변환하는 신경망 구조를 제안하였으며, 해당 구조는 이후의 여러 연구들의 기반이 되었다 [2, 4, 5]. Faceformer [4]는 얼굴 애니메이션 생성을 위해 자기 회귀 모델 중 하나인 트랜스포머 모델 구조를 도입함으로써 정확도를 크게 향상시켰다. CodeTalker [5]는 음성 신호와 표정 사이의 다중 매핑 문제를 해결하기 위해 얼굴 동작에 대한 잠재코드북 (latent codebook)을 도

입하였다. Imitator [6]는 참조 비디오에 나타나는 화자 고유의 특징 임베딩을 추정하고 최적화하는 방식을 통해 비디오 속 화자의 발화 방식을 효과적으로 포착하여 얼굴 애니메이션을 생성한다.

이러한 방법들은 정확도의 개선과 생성 애니메이션의 다양성 측면에서 높은 성과를 보였으나, 공통적으로 학습에 사용된 얼굴 메쉬 구조에 한정하여 적용 가능하다는 한계가 있다. 또한, 학습 설정에 민감하여 새로운 얼굴 메쉬에 적용할 경우 세심한 파라미터 조정이 필요하기에, 범용성이 다소 제한적이다. 이에 반해, 본 연구에서 제안하는 방법은 메쉬 변형 네트워크를 활용하여 임의의 메쉬 구에 적용할 수 있는 메쉬 구조에 비종속적(mesh-agnostic) 구조를 지니고 있어, 음성 기반 얼굴 애니메이션 생성 시 높은 확장성과 범용성을 제공한다.

2.2 음성 기반 파라매터 추정

앞서 언급한 정점 예측 기반 방법들 [1, 2, 3, 4, 5]과 달리, 음성 신호로부터 얼굴 발화 애니메이션을 생성하기 위해 리깅 파라미터나 블랜드쉐입 (Blendshape) 계수를 예측하는 연구도 활발히 진행되어 왔다. VisemeNet [15]은 LSTM 계층을 이용해 2D JALI viseme 필드 파라미터를 예측하며, JALI [16] viseme을 기반으로 리깅된 캐릭터 모델에 한하여 적용할 수 있다. Voice2Face [17]는 조건부 변분 오토인코더 (CVAE)를 사용해 음성 신호로부터 얼굴 메쉬 정점 위치값을 직접 복원하고, 이어서 MLP를 통해 복원된 정점으로부터 리깅 파라미터를 회귀한다. 또한, Medina 와 그의 동료들 [18]은 사실적인 발음 애니메이션에서 중요한 요소인 혀 움직임을 고려하여 입술과 혀의 랜드마크를 예측하고, 이를 바탕으로 혀 모델 구축 및 혀 애니메이션 생성을 위한 리깅 파라미터 추정 방법을 제안했다. 본 연구에서는 이와 유사한 맥락에서, 입력 음성 신호로부터 FACS 기반 블랜드쉐입 파라미터에 대응하는 해석 가능한 표정 잠재코드를 예측한다. 이렇게 얻어진 잠재코드는 사전학습한 디코더를 통해 대상 얼굴 모델을 매 프레임 마다 변형함으로써 얼굴 발화 애니메이션을 생성한다.

2.3 Deformation Transfer

변형 전달 (deformation transfer)은 한 메쉬에서의 변형 정보를 다른 메쉬로 전달하는 작업이다. 초기 연구에서는 정점 단위의 변위나 변형 그래디언트를 전달하기 위해 소스 메쉬와 타겟 메쉬 간의 대응점을 수동으로 설정해야 했다 [19, 20]. 이후 심화학습 기반 신경망의 등장으로 일부 연구 [21, 22]는 이를 신경망의 잠재 공간 (latent space)을 매개하는 방식을 통해 대응점 지정 문제를 극복하였다. 그러나 이러한 접근 방식들 역시 학습한 메쉬 구조에 종속되므로, 임의의 메쉬에 대해 곧바로 적용하기 어렵다는 한계가 있다. 또 다른 접근으로는, 두 얼굴 메쉬 간의 표정 전이를 위해 이미지-투-이미지 (image-to-image) 변환 네트워크를 활용하는 방법들이 제안되었다 [23, 24]. 하지만 이 경우에도 새로운 얼굴 모델마다 별도의 학습 과정이 필요하다는 한계가 있다. 최근, Aigerman 과 그의 동료들 [25]은 삼각형 단위의 야코비안 (Jacobian)을 활용하여 대응점의 지정 없이 임의의 메쉬 쌍 간에도 deformation transfer가 가능한 방식을 제안하였다. 이어서 NFR [7]은 주어진 얼굴 모델에서의 전역적인 형태와 표정 정보를 추출하는 인코더를 추가함으로써 메쉬 구조에 구애받지 않는 리타게팅이 가능하도록 확장하였다. 본 연구는 NFR의 접근법에 기초하여, 음성 기반의 3차원 얼굴 발화 애니메이션이라는 영역으로 해당 방법론을 확장한다.

3 알고리즘

본 연구의 목적은 주어진 입력 음성 신호를 기반으로 임의의 메쉬 구조의 3D 얼굴 모델에 대한 발화 애니메이션을 생성하는 것이다. 이를 위해 본 연구는 두 가지 핵심 구성 요소를 활용한다: (1) 음성 신호를 해석 가능한 FACS 기반 표정 코드로 변환하는 매핑 네트워크인 *Wav2Rig*, 그리고 (2) 변환된 표정 코드에 따라 대상 얼굴 메쉬 M을 변형시키는 NFR [7] 기반의 메쉬 변형 네트워크이다. 이 문제는 다음과 같이 수식으로 정의된다:

$$\hat{M}^{1:t} = \phi(s, a, M), \tag{1}$$

여기서 s는 스타일 임베딩, a는 입력 음성 신호, $\hat{M}^{1:t}$는 길이 t의 애니메이션 시퀀스를 나타낸다. 함수 ϕ는 wav2vec 2.0 [8], Wav2Rig, NFR의 인코더 및 디코더를 포함한 전체 시스템을 포괄한다. 제안하는 방법의 전체 구조는 그림 2에 제시되어 있으며, 이후의 절에서는 NFR의 개요(Sec. 3.1), Wav2Rig의 세부 구성(Sec. 3.2)과 추론단계 (Sec. 3.3), 그리고 데이터셋 구성 방법(Sec. 3.4)에 대해 구체적으로 설명한다.

3.1 NFR 알고리즘

본 절에서는 제안하는 연구의 기반이 되는 사전학습 모델 NFR [7]의 알고리즘에 대해 설명한다. NFR은 서로 다른 구조를 가진 메쉬 간에도 변형 (deformation)을 리타게팅할 수 있도록 설계된 신경망 기반의 방법이다. NFR은 Identity 인코더, Expression 인코더 그리고 디코더로 구성된다. Identity 인코더는 대상 메쉬로부터 identity 잠재코드 $z_i \in \mathbb{R}^{100}$를 추정하며, Expression 인코더는 소스 메쉬로부터 expression 잠재코드 $z_e \in \mathbb{R}^{128}$를 추정한다. 이렇게 얻어진 코드들 (z_i, z_e)과 함께, 이미지 코드 $c \in \mathbb{R}^{128}$, 그리고 대상 메쉬의 삼각형에 해당하는 중심점 (centroid) 및 법선 벡터 (normal vector)로 구성된 각 삼각형별 피처 코드 $\beta_j \in \mathbb{R}^6$를 디코더 Ψ에 입력하면, 대상 메쉬의 j번째 삼각형 γ_j에 대한 변형 행렬 $P_j \in \mathbb{R}^{3\times3}$을 생성할 수 있다. 이 과정은 다음과 같이 수식으로 표현된다:

$$P_j = \Psi(\beta_j, z_i, z_e, c). \tag{2}$$

각 삼각형 γ_j에 대해 P_j를 예측한 이후, 최종적으로 변형된 메쉬의 정점들을 복원하기 위한 추가 단계가 필요하다. 먼저 예측된 변형 행렬 P_j를 해당 삼각형의 tangent basis에 제한하여 $R_j \in \mathbb{R}^{3\times2}$로 변환하고, 푸아송 (Poisson) 방정식을 통해 최종 정점 위치 V^*를 얻는다. 해당 과정은 다음과 같은 최적화 문제로 표현된다:

$$V^* = \min_V \sum |\gamma_j| \left\| V\nabla_j^\top - R_j \right\|^2, \tag{3}$$

여기서 $V\nabla_j^\top$는 삼각형 γ_j의 야코비안(Jacobian)을 나타내며, ∇_j는 대상 메쉬 정점 V를 해당 삼각형의 야코비안으로 매핑하는 그래디언트 연산자이고, $|\gamma_j|$는 삼각형의 면적을 의미한다.

NFR의 강점은 해석 가능한 표현 잠재 공간을 제공하면서, 특정한 메쉬 구조나 형태에 제한되지 않는다는 점에 있다. 표정 잠재코드 z_e의 앞 53차원은 ICT-FaceKit [10]에서 제공하는 FACS 기반 블랜드쉐입 파라미터와 동일한 방식으로 동작하기에 LiveLinkFace [26] 앱과 같은 상용 프로그램과의 호환이 용이하다. 본 연구에선 이 53차원을 z_{FACS}로, 나머지 75차원은 z_{ext}로 구분하여 표기한다.

음성 신호가 정확히 z_e로 매핑될 수 있다면, 사전학습한 디코더 Ψ를 통해 별도의 추가 학습이나 리타게팅 절차 없이도 입력된 음성에 맞춰 어떤 3D 얼굴 메쉬라도 발화 애니메이션을 생성할 수 있다. 이러한 가능성을 바탕으로, 본 연구에서는 음성 신호를 z_e로 직접 변환할 수 있는 새로운 인코더 *Wav2Rig*을 설계한다.

3.2 음성 신호 매핑 네트워크

*Wav2Rig*은 사전학습 음성 신호 분석 모델의 딥피처(deep feature)를 사전학습 변형 전달 모델의 잠재코드 z_e로 변환하는 매핑 네트워크이다. 사전학습 음성 신호 분석 모델로는 wav2vec2.0 [8]을 활용하며, 이를 통해 a 로부터 딥피처 f 를 추출한다. ($a \rightarrow f^{1:t}, f^{1:t} \in \mathbb{R}^{t\times768}$). wav2vec2.0은 컨볼루션 기반의 전처리 네트워크와 트랜스포머 인코더(transformer encoder)로 구성되어 있다. 트랜스포머 인코더는 총 12개의 계층으로 이루어져 있으며, 각 계층은 서로 다른 수준의 표현 정보를 담고 있기 때문에, 달성하고자 하는 세부 목적에 따라 가장 적합한 계층이 달라질 수 있다[27]. 기존 방법들 [4, 5]은 wav2vec 2.0의 파라미터를 파인튜닝하고 마지막 계층에서 추출한 피처를 사용한다. 반면, 본 방법은 wav2vec 2.0 파라미터를 고정한 후, 트랜스포머 인코더의 중간 계층에서 추출된 피처 $f^{1:t}$만을 사용한다. 계층 별 피처의 효과는 Sec. 4.4.1에서 자세히 분석하였으며, 해당 실험을 통해 중간 계층이 가장 효과적임을 확인하였다. 딥피처 $f^{1:t}$와 발화 스타일 임베딩 $s \in \mathbb{R}^{128}$은 *Wav2Rig*에 입력되어 표정 코드 시퀀스 $\hat{z}_e^{1:t}$를 예측한다. 이 과정은 다음과 같이 표현된다:

$$\hat{z}_e^{1:t} = \text{Wav2Rig}(f^{1:t}, s). \tag{4}$$

네트워크 구조는 입력 계층, 두 개의 은닉 계층, 그리고 출력 계층으로 구성된 총 네 개의 1D 컨볼루션(Convolution) 계층으로

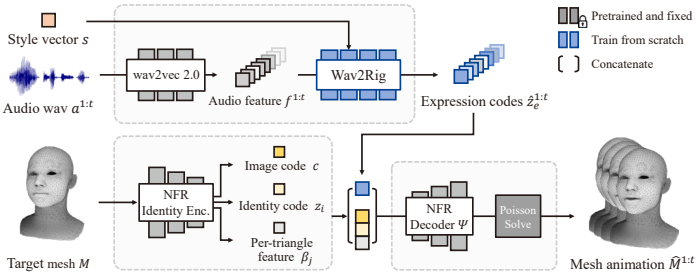


Figure 2: **Overview.** Given raw audio and a target 3D face mesh as input, our method produces audio-synchronized facial animation for the target 3D face mesh.

이루어져 있다. 입력 계층은 오디오 피처 $f^{1:t}$을 받아 이를 처리한 뒤 은닉 계층에 전달한다. 이때 은닉 계층의 출력은 발화 스타일 임베딩 s와 함께 연결(concatenate)된다.

본 설계는 CodeTalker [5]에서 제안된 방법과 유사하게 one-hot 라벨 대신 학습 가능한 스타일 벡터를 활용함으로써 네트워크가 스타일 공간을 내재적으로 학습할 수 있도록 유도한다. 스타일 벡터는 데이터셋의 각 화자에 대해 고유하게 할당하여 학습에 사용되며, $s_n \in S, S = s_1, s_2, \ldots, s_N$로 정의된다. 이때 N은 전체 화자 수를 의미한다. 은닉 계층의 출력을 기반으로, 최종 출력 계층은 표정 코드 시퀀스 $\hat{z}_e^{1:t}$을 생성한다.

*Wav2Rig*의 학습은 손실함수 L_{rig}만을 사용한다. L_{rig}는 정답값과 예측된 표정 코드 간의 L1 거리를 측정하며, 다음과 같이 정의된다:

$$L_{rig} = \left\| z_e^{1:t} - \hat{z}_e^{1:t} \right\|_1, \tag{5}$$

여기서 $z_e^{1:t}$는 정답 블랜드쉐입 파라미터인 $z_{\text{FACS}}^{\text{GT}}$와 z_{ext}^0로 구성된 벡터 시퀀스를 의미한다. 시퀀스의 각 시점에서의 표정 코드 z_e는 다음과 같이 정의된다:

$$z_{e_k} = \begin{cases} z_{\text{FACS}_k}^{\text{GT}} & \text{if } k \leq 53 \\ z_{\text{ext}}^0 & \text{if } k > 53, \end{cases} \tag{6}$$

여기서 k는 표정 잠재코드의 인덱스를 나타내며, z_{ext}^0는 모든 요소가 0으로 채워진 제로 벡터이다.

3.3 추론단계

본 방법은 추론 단계에서 완전한 엔드투엔드(end-to-end) 파이프라인으로 동작하며, Eq. (1)에 따라 음성신호, 대상 얼굴 모델, 그리고 스타일 임베딩을 입력으로 받아 발화 애니메이션을 생성한다. 추론과정은 다음과 같다. 먼저, 음성 신호로부터 추정한 wav2vec2.0의 피처를 입력으로 Wav2Rig은 표정 잠재코드 시퀀스를 추정한다. 다음으로 사전학습 NFR의 identity 인코더는 대상 얼굴 모델로부터 z_i, 이미지 코드 c, 그리고 삼각형 단위의 피처 코드(per-triangle feature code) β_j 를 추출한다. 이렇게 얻어진 코

드들은 결합되어 디코더 Ψ에 입력되며, 이를 통해 삼각형 단위의 디포메이션 필드 시퀀스가 생성된다. 이후, Eq. (3)을 통해 최종적으로 음성에 동기화된 3D 얼굴 메쉬의 애니메이션이 복원된다. 전체적인 파이프라인 구성은 Fig. 2에 제시되어 있다.

3.4 데이터

Wav2Rig 학습하기 위해, 음성 파일 (a)과 FACS 기반 블랜드쉐입 파라미터 ($z_{\text{FACS}}^{\text{GT}}$)로 구성된 데이터셋을 구축하였다. 블랜드쉐입 파라미터는 iPhone 13 Pro를 삼각대에 고정한 상태에서 LiveLinkFace [26] 앱을 사용하여 수집하였으며, 해당 앱은 영상으로부터 음성과 함께 FACS 기반의 ARkit 블랜드쉐입 파라미터를 제공한다. 획득한 ARkit 블랜드쉐입 파라미터는 ICT-FaceKit [10]이 제공하는 블랜드쉐입 매핑에 따라 $z_{\text{FACS}}^{\text{GT}}$로 변환되어 저장되었다. 영상은 초당 30프레임으로 녹화되었고, 음성은 44,100 Hz의 샘플링 레이트로 수집되었다. 데이터 수집에는 총 16명 (남성 7명, 여성 9명)이 참여하였으며, 각 참가자에 대해 71개의 문장에 관한 발화 데이터를 확보하였다. 각 화자에 대해서는 ICT의 Identity 블랜드쉐입에서 임의의 얼굴 형태를 샘플하여 각 참자가 별 발화 애니메이션에 할당하였다. 실험에 사용한 모든 얼굴 모델은 NFR에서 수행된 메쉬 표준화 절차에 따라 전처리되었으며, 이 과정에서 눈구멍과 입 안쪽 내부는 제외하였다. 전체 데이터는 학습용 56개, 테스트용 15개로 분할하여 실험에 사용하였다.

3.5 구현 세부사항

오디오 피처는 별도의 언급이 없는 경우, wav2vec 2.0 [8]의 5번째 계층에서 추출한 것을 사용하였다. Wav2Rig은 Adam optimizer [28]를 사용하여 학습하였으며, 하이퍼파라미터는 $\beta_1 = 0.5$, $\beta_2 = 0.999$, 학습률은 1×10^{-4}로 설정하였다. 학습은 총 200에폭(epoch)에 걸쳐 진행되었으며, 오디오 피처는 윈도우 크기 8로 슬라이싱하여 입력하였다. 모든 학습 및 실험은 단일 NVIDIA RTX A5000 GPU에서 수행되었으며, Wav2Rig의 학습에는 약 1.5시간이 소요되었다.

4 실험

4.1 성능 평가지표

본 방법의 성능을 평가하기 위해, Sec. 3.4에서 설명한 ICT-FaceKit [10] 기반의 테스트셋을 사용하였다. 추가적으로, 다양한 메쉬 구조에 대한 일반화 성능을 검증하기 위해 Multiface [11], VOCASET [12], BIWI [13] 등 여러 공개 데이터셋으로부터 확보한 얼굴 모델을 활용하였다. 해당 얼굴 모델은 Fig. 3와 같이 메쉬 표준화 절차에 따라 전처리된 후 사용되었다. 평가지표로는 기존 연구들 [4, 5]에서 사용된 Lip Vertex Error (LVE)를 채택하였다. LVE는 각 프레임마다 예측된 입술 정점과 정답 정점 간의 L2 거리 중 최대값의 평균을 측정한 값이다. 입술 정점은 ICT-FaceKit 프로젝트에서 사전에 지정된 랜드마크를 사용하였다 [1].

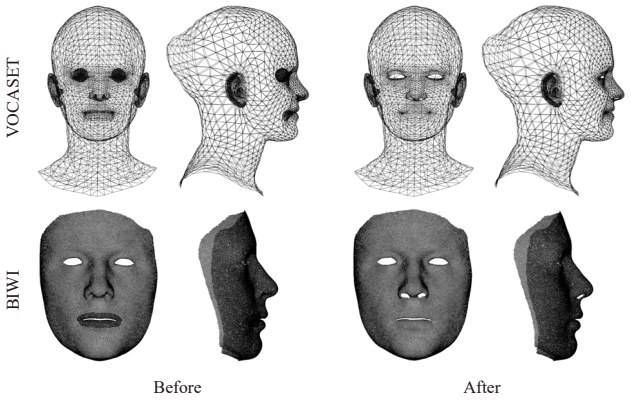


Figure 3: **Mesh standardization.** Eye sockets and mouth internals are removed from VOCASET [12] meshes, and both mouth and nose internals are removed from BIWI [13] meshes.

4.2 임의 메쉬 구조 대응 실험

본 실험의 목표는 학습한 메쉬 구조와 다른 구조를 가진 얼굴 모델에 관한 대응 여부 및 성능을 평가하는 데 있으며, 시각적 결과는 Fig. 4에 제시되어 있다. 학습에 사용된 ICT [10] 기반의 얼굴 모델에 대해서는 물론, 학습 데이터에 포함되지 않은 얼굴 모델 [11, 12, 13]에 대해서도 본 방법으로 발화 애니메이션을 강건하게 생성할 수 있는 것을 확인할 수 있었다.

본 방법의 성능을 정량적으로 평가하기 위해, Sec. 3.4에서 기술한 테스트 데이터를 decimation과 subdivision을 통해 리메싱 (re-meshing)하여 다양한 메쉬 구조를 갖는 정답 애니메이션을 확보한 뒤, 생성된 애니메이션 결과와의 LVE를 측정하였다. decimation이 적용된 메쉬의 경우, 먼저 Blender [2]를 사용하여 메쉬의 정점 수를 줄였다. 이후, 원본 메쉬와 decimation이 적용된 메쉬 간의 정점 대응 관계를 설정하고, 각 프레임에서 원본 메쉬의 변형 결과를 따라가도록 decimation이 적용된 메쉬를 변형하여 발화 애니메이션을 생성하였다. subdivision이 적용된 메쉬의 경우, Loop subdivision [29]을 적용하여 해상도를 높인 뒤, decimation이 적용된 메쉬와 동일한 방식으로 대응점 설정을 통해 발화 애니메이션을 생성하였다. 평가를 위한 입술 랜드마크의 경우, 원본 메쉬의 입술 랜드마크와 가장 가까운 정점을 기준으로 대응 정점을 선택하였다.

정량적 결과는 Tab. 1에 제시되어 있다. decimation이 적용된 메쉬 (약 5천 개의 삼각형)는 학습에 사용된 원래의 메쉬 구조 (약 1만5천 개의 삼각형)와 유사한 성능을 보였다. subdivision이 적용된 메쉬 (약 6만2천 개의 삼각형)의 경우, 원본 메쉬 구조를 기반으로 생성한 애니메이션과의 유사한 오차 범위를 보여주었다.

[1] https://github.com/ICT-VGL/ICT-FaceKit
[2] https://www.blender.org/

Table 1: Comparison of performance achieved by various triangulation on ICT face model.

Triangulation	LVE ↓ ($\times 10^{-3}$)
Original (15K faces)	1.1776 ± 0.7796
Decimated (5K faces)	1.1484 ± 0.8808
Loop subdivided (62K faces)	1.3742 ± 0.9870

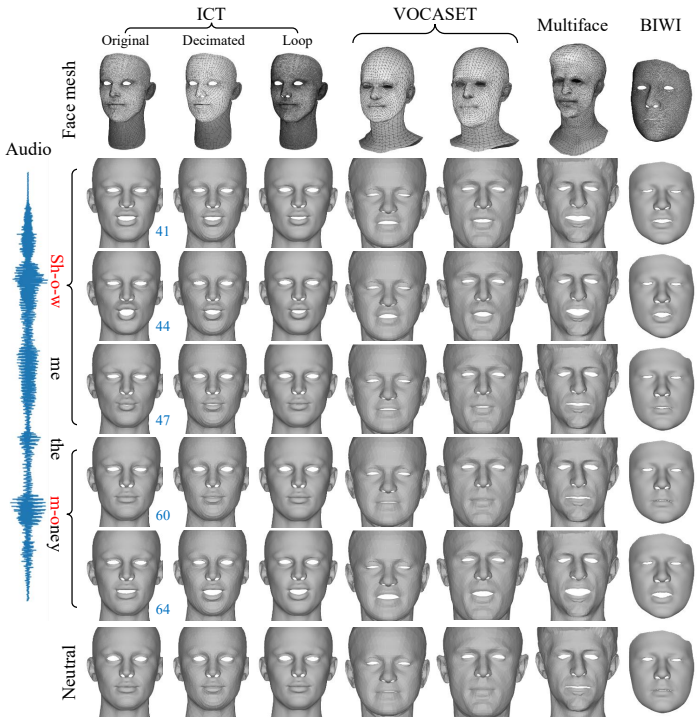


Figure 4: **Animated 3D faces with varying shapes and mesh structures.** Visual results are shown across different mesh typologies including ICT (in-domain) [10], VOCASET [12], BIWI [13] and Multiface [11]. Blue numbers indicate the frame indices of the generated animations.

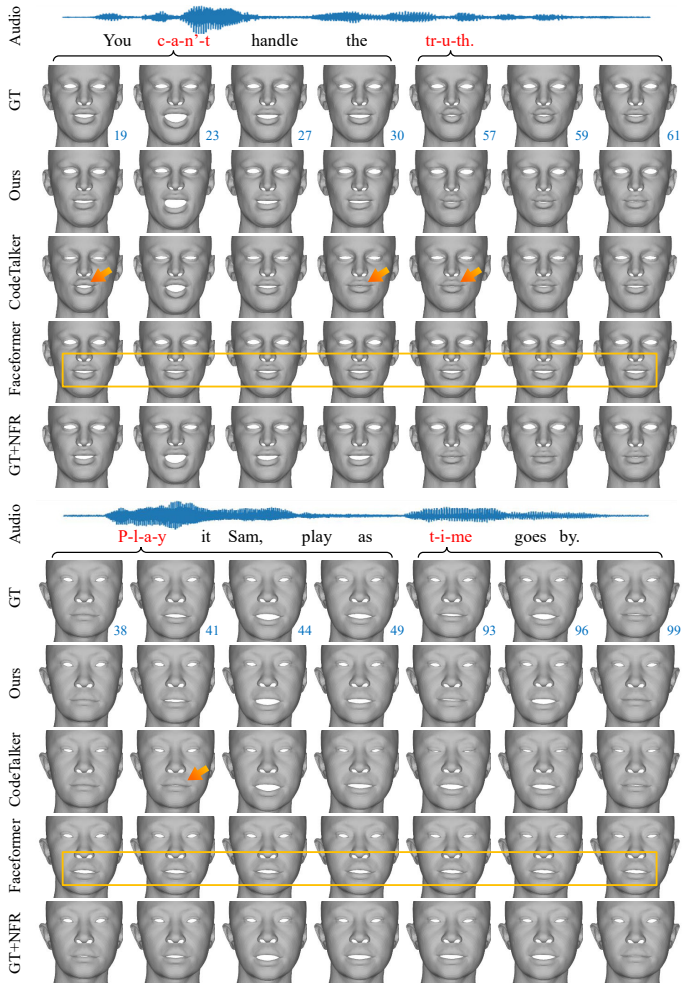


Figure 5: **Visual comparison with competing methods.** Blue numbers indicate the frame indices of the generated animations. Orange arrows indicate incorrect lip-sync of CodeTalker [5]. Yellow box indicates static motion of Faceformer [4].

4.3 기존 방법과의 비교

정량적 성능 평가를 위해 음성 기반 얼굴 애니메이션 기법인 Faceformer [4]와 Codetalker [5]와의 비교 실험을 진행하였다. 두 방법은 본 방법과 달리 학습된 메쉬 구조에 종속된 mesh-specific 방식으로, 정점 위치 또는 변위를 직접 예측하여 얼굴 애니메이션을 생성한다. 공정한 비교를 위해 각 방법은 공식 구현 코드를 기반으로 Sec. 3.4에서 구축한 ICT 데이터를 사용하여 처음부터 재학습하였다. 학습이 완료된 후, 동일한 테스트셋을 사용하여 애니메이션을 시각화하고 LVE를 측정하였다. 또한, *Wav2Rig*의 예측이 완벽할 경우 얻을 수 있는 성능 상한을 파악하기 위해, 정답 FACS 기반 블랜드쉐입 파라미터와 NFR을 조합한 결과($z^{\text{GT}}_{\text{FACS}}$ + NFR)도 함께 비교하였다. 공개 데이터셋인 VOCASET [12]과 BIWI [13]의 애니메이션 데이터는 음성 오디오-얼굴 애니메이션 쌍만 제공하기에 본 연구의 학습 조건 (음성 오디오와 블랜드쉐입 파라미터 쌍) 과 맞지 않아 비교에서 제외하였다.

정량적 결과는 Tab. 2 에서 확인할 수 있다. Codetalker [5]는 LVE에서 가장 안 좋은 결과를 보였으며, 이는 발화 애니메이션과 입력 음성 신호 간 동기화 오류가 자주 발생했기 때문으로, Fig. 5의 주황색 화살표로 확인할 수 있다. Faceformer는 Codetalker보다 개선된 LVE 값을 보였지만, 결과 애니메이션이 다소 정적(static)으로 나타났으며, 이는 평균적인 표정으로 수렴한 결과로 해석된다. 해당 특징은 Fig. 5의 노란색 박스로 강조되었다. 반면, 제안하는 방법은 가장 낮은 LVE 값을 기록하였으며, 이는 상한선($z^{\text{GT}}_{\text{FACS}}$ + NFR)에 근접한 결과를 보여준다. 특히 단어 "*truth*"를 발음하는 Fig. 5의 프레임(57–61) 구간에서, 본 방법은 입술 변형이 발음에 걸맞게 표현된 것을 확인할 수 있다.

4.4 설계 검증

본 절에서는 제안하는 *Wav2Rig*의 설계의 효율성을 검증하기 위해 다음 세 가지 핵심 요소를 분석한다: (1) wav2vec 2.0 [8]에서 추출한 계층별 오디오 피처에 따른 성능 변화, (2) Wav2Rig 네트

Table 2: **Comparison with audio-driven facial animation methods.** The best LVE value is marked in **bold**. "$z_{\text{FACS}}^{\text{GT}}$+NFR" indicates the upper bound of the performance of our method.

Method	Mesh-agnostic	LVE ↓ ($\times 10^{-3}$)
CodeTalker	✗	1.5927 ± 0.8608
Faceformer	✗	1.4854 ± 0.9858
Ours	✓	**1.1776** ± 0.7796
$z_{\text{FACS}}^{\text{GT}}$ + NFR	-	1.0642 ± 0.7425

워크 구조의 효율성, (3) z_e의 변주에 따른 성능 변화. 다음 절에서 각 요소에 관한 실험과 결과를 통해 선택의 정당성을 입증한다.

4.4.1 wav2vec 2.0의 계층별 오디오 피처 분석

본 실험에서는 wav2vec 2.0 [8]의 다양한 계층에서 추출된 오디오 피처를 학습에 사용하여, 가장 효과적인 피처 계층을 확인하였다. Sec. 3.4에서 구축한 데이터셋을 사용하여 실험을 진행했으며, 각 계층에서 추출한 피처로 계산한 LVE 값을 Fig. 6에 시각화하였다. 그 결과, 중간 계층(5 ∼ 9층)의 피처가 가장 우수한 성능을 보였으며, 특히 5번째 층의 피처에서 최저 LVE를 기록하여 본 방법의 기본 구성으로 채택하였다. 이와 같은 중간 계층 피처의 우수한 성능은, wav2vec 2.0이 입력의 일부를 가린 후 이를 예측하는 방식으로 학습된 데에서 기인한 것으로 보인다. 이 학습 전략은 트랜스포머의 중간 계층에 고수준의 의미 정보(semantics)를, 마지막 계층에는 저수준의 음향 정보(sound detail)를 내장하도록 유도한다. 따라서 wav2vec 2.0의 중간 계층을 활용하는 것은 본 연구뿐 아니라, CodeTalker [5]나 Faceformer [4]와 같은 메쉬 구조에 종속적인 방법에서도 성능 향상을 가져올 수 있을 것으로 사료된다.

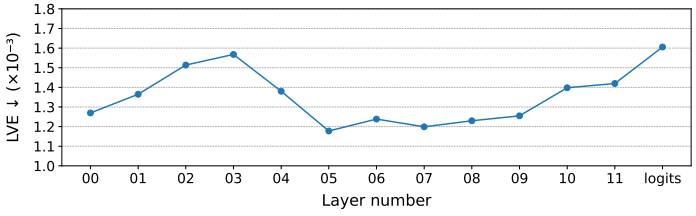


Figure 6: **Comparison of lip vertex errors for the features from different layers of wav2vec 2.0 [8].**

4.4.2 Wav2Rig 네트워크 구조

*Wav2Rig*은 오디오 피처를 입력으로 받아 이에 대응하는 표정 잠재코드를 출력하는 합성곱 신경망 (CNN) 기반 모델이다. 기존의 연구들은 트랜스포머 계층 [30]이나 VQ-VAE [31]와 같은 복잡한 구조를 활용하였다. 본 절에서는 오디오-표정 잠재코드 예측 과제에 있어 단순한 CNN 구조로도 충분한지를 검증하고자 한다.

이를 위해, 기존 방법들의 네트워크 구조를 참고하여 표정 코드를 직접 예측할 수 있도록 일부를 수정한 뒤, 동일한 실험 조건에서 성능을 비교하였다. 공정한 비교를 위해 모든 네트워크는 wav2vec 2.0의 파라미터를 고정한 상태로 Sec. 3.4의 학습 데이터셋을 사용해 처음부터 학습하였으며, 모든 구조에 스타일 임베딩도 동일하게 포함시켰다. 실험 결과는 Tab. 3에 제시되어 있으며, Faceformer 및 CodeTalker 기반 구조는 본 연구의 CNN 기반 모델보다 낮은 성능을 보였다. 이는 오디오-표정 코드 매핑 과제에 있어 CNN 구조가 보다 적합하다는 점을 시사한다.

Table 3: **Performance of various network architectures.** The best score is marked in **bold**.

Architecture	LVE ↓ ($\times 10^{-3}$)
Faceformer	1.6256 ± 0.5127
CodeTalker	2.0969 ± 0.8608
Ours	**1.1776** ± 0.7796

Table 4: **The impact of our design choices.** The best score is marked in **bold**.

Style	z_e variant	LVE ↓ ($\times 10^{-3}$)
✗	case A	1.9249 ± 0.5930
	case B	1.8868 ± 0.5694
	case C	2.8791 ± 0.7581
✓	case A	**1.1776** $\pm$ **0.7796**
	case B	1.2519 ± 0.4224
	case C	2.5165 ± 0.5695

4.4.3 z_e 변주에 따른 성능 변화

본 실험에서는 학습과정에서 z_{ext}의 변주에 따른 모델의 성능 변화를 확인하고자 한다. NFR의 표정 잠재코드는 FACS 기반 해석 가능한 잠재코드와 확장코드로 구성된다 $z_e^{\text{NFR}} = [z_{\text{FACS}}^{\text{NFR}}, z_{\text{ext}}^{\text{NFR}}]$. 이 때 확장코드 $z_{\text{ext}}^{\text{NFR}}$는 학습 과정에서 암시적으로 (implicit) 학습되며, 학습 데이터에서 블랜드쉐입으로 표현되지 않는 기타 움직임 정보를 포함한다. 이를 바탕으로 *Wav2Rig* 학습 과정에서 다음과 같은 세 가지 z_e 설정을 사용하여 비교하였다.

첫 번째 설정은 Eq. (6)를 그대로 사용하는 방식으로, 본 방법의 기본 구성이다. 두 번째 설정은 사전학습한 NFR로 추정한 표정 잠재코드를 그대로 사용하는 방식이다. 세 번째 설정은 정답 FACS 기반 블랜드쉐입 파라미터 z_{FACS}와 사전학습한 NFR로 추정한 확장 코드 $z_{\text{ext}}^{\text{NFR}}$를 결합하는 방식이다:

$$z_e = \begin{cases} [z_{\text{FACS}}^{\text{GT}}, z_{\text{ext}}^{0}] & \text{if case A} \\ [z_{\text{FACS}}^{\text{GT}}, z_{\text{ext}}^{\text{NFR}}] & \text{if case B} \\ [z_{\text{FACS}}^{\text{NFR}}, z_{\text{ext}}^{\text{NFR}}] & \text{if case C} \end{cases} \tag{7}$$

세 가지 변형을 적용한 결과는 Tab. 4에 제시되어 있다. case A와 case B는 유사한 성능을 보인 반면, case C의 경우 성능이 저하되었다. 이는 z_e^{NFR}에 포함된 NFR의 예측 오차의 답습과 학습 데

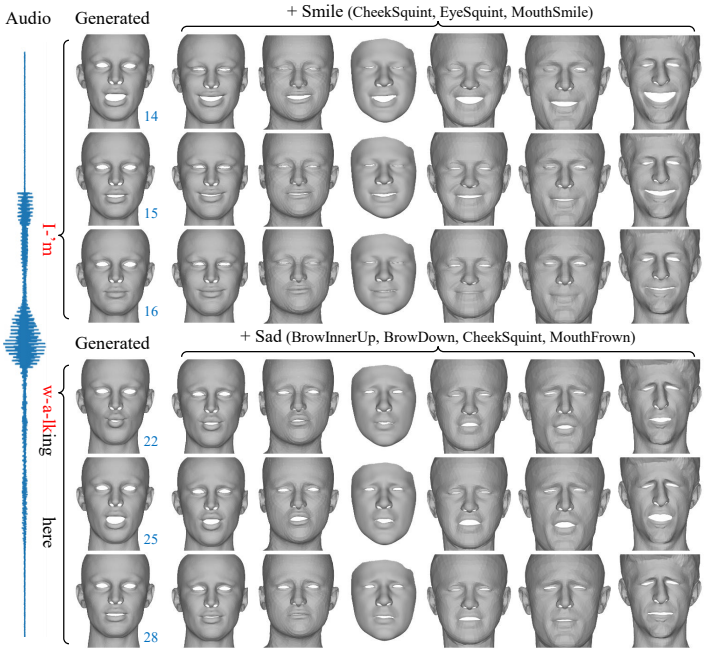


Figure 7: **Edited result of the generated speech animation.** Visual results are shown across different mesh structures, marked with specific expressions and the related facial rig names. From the left is the generated animation in ICT [10], followed by edited meshes, two from ICT (in-domain), one BIWI [13], two VOCASET [12], and one Multiface [11]. Blue numbers indicate frame indices of the generated animations.

이터의 bias로 인해 FACS 기반 잠재코드와 확장 코드 간의 얽힘 (entanglement)을 학습했기 때문인 것으로 추정된다. 결과적으로 정확한 FACS 기반의 표정 잠재코드 학습이 본 방법에서 결정적으로 작용함을 시사한다. 또한, 모든 설정에서 스타일 임베딩을 추가했을 때 성능이 크게 향상되었다. 이러한 실험 결과를 바탕으로, 본 연구는 Eq. (6)를 기본으로 하되 스타일 임베딩을 보완하여 사용하는 방안을 채택하였다.

5 응용 사례

5.1 표정 편집 및 후처리

기존 연구들 [3, 4, 5]은 대부분 입력 음성으로부터 얼굴 움직임을 직접적으로 예측하는 데 초점을 두고 있으며, 이후 생성된 애니메이션을 수정하거나 편집하는 기능은 고려하지 않았다. 반면, 본 연구의 *Wav2Rig*은 음성 신호를 FACS 기반 해석 가능한 표정 잠재코드로 매핑하기 때문에, 블랜드쉐입 애니메이션과 마찬가지로 특정 표정을 선택적으로 조정하거나 편집할 수 있는 유연함을 제공한다. 특히, 해석 가능한 표정 잠재코드는 ICT-FaceKit [10]에서 제공하는 블랜드쉐입과 연동되어, 직관적인 편집 환경 제공이 가능하다. Fig. 7에서 생성된 애니메이션에 'smile' 또는 'sad' 등의 감정에 맞게 편집된 사례를 확인할 수 있다.

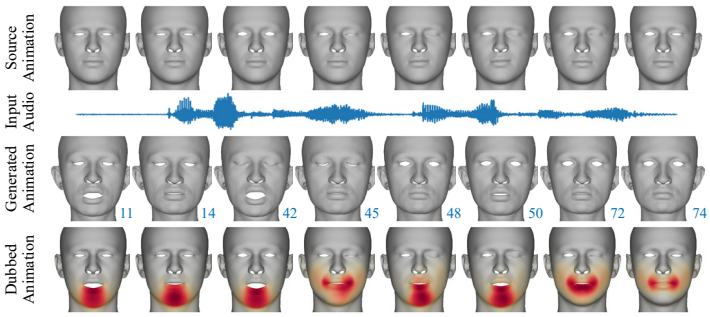


Figure 8: **Visual dubbing result.** The vertices are colored by the L2 distance between the meshes from the source animation and the dubbed animation. Blue numbers indicate frame indices of the generated animations and the dubbed animations.

5.2 더빙

더빙은 본 방법이 적용될 수 있는 또 다른 응용 사례이다. 본 방법은 입력 음성으로부터 주어진 얼굴 모델에 적절한 발화 애니메이션을 생성함과 동시에, 원본 애니메이션에서의 상안면 움직임은 유지할 수 있다. Fig. 8의 세 번째 행에서 보이는 바와 같이, 입 주변 영역에서 높은 L2 거리 값이 뚜렷하게 나타나 발화 애니메이션이 적용된 것을 확인할 수 있으며, 상안면 영역에선 L2 거리 값이 전무하나 경계에선 부드럽게 보간 (interpolate) 되어 원본 애니메이션을 해치지 않고 자연스럽게 통합됨을 확인할 수 있다.

6 토론

본 방법의 성능 평가에 있어 정량적 지표 LVE를 활용하였으며, 기존 방법 대비 우수한 성능을 보여주었다. 해당 지표는 생성한 발화 애니메이션의 입술에 대한 최대 오차의 평균으로 생성한 애니메이션의 자연스러움과 같은 정성적 평가를 포괄한 완전한 지표로 보기는 어렵다. 이에 차후 사전학습 모델 기반 perceptual 지표 또는 사용자 평가를 통한 주관적 지표를 활용하여 본 방법의 정성적 측면의 평가를 보강하고자 한다. 또한 본 방법은 CNN를 활용하여 기존 방법 대비 연산적 측면에서 이점을 갖고 있다. 이에 향후 실제 어느정도의 연산량 개선이 가능한지와 이를 통한 응용 가능성을 탐구해볼 필요가 있다.

7 한계점

본 방법은 임의의 메쉬 구조에 대해 음성 기반 발화 애니메이션 생성이 가능하다는 장점을 갖지만, 몇 가지 한계도 존재한다. 첫째, 사전학습 모델 NFR의 설계 특성상 최적의 성능을 위해 대상 메쉬가 정렬 (aligned) 되고 표준화 (standardized) 되어야 한다. 특히 눈구멍이나 입 안쪽 등에서 메쉬 표준화가 충분히 이루어지지 않은 경우, 변형 전달 (deformation transfer)에 의존하는 구조로 인해 부자연스러운 움직임이 발생하는 경우가 관찰되었다. 둘째, 본 방법은 사람 얼굴 메쉬를 기반으로 학습되었기 때문에, 사람

과 유사한 비율을 가진 얼굴에만 적용이 가능하다. 과장된 비율 가진 캐릭터 얼굴의 경우, 원하는 품질의 결과를 얻기 어렵다. 마지막으로, 본 방법은 시간적 일관성 (temporal coherency)을 명시적으로 보장하지 않기 때문에, 프레임 간 메쉬 변형이 일관되지 않게 예측될 수 있으며, 이로 인해 생성된 애니메이션에서 떨림 (jitter)이 발생할 수 있다.

8 결론 및 향후 연구

본 연구에서는 메쉬 구조에 구애받지 않는 음성 기반 3차원 얼굴 발화 애니메이션 생성 방법을 제안하였다. 본 방법은 입력 음성 신호를 FACS 기반의 해석 가능한 표정 잠재코드로 매핑하고, 메쉬 구조에 비종속적인 사전학습 변형 전달 모델 NFR을 활용하여 임의의 메쉬 구조 및 형태를 가진 얼굴 모델에 적용 가능하다. 본 시스템의 핵심인 *Wav2Rig*는 사전학습한 wav2vec 2.0의 5번째 계층 피처를 사전학습 모델 NFR의 표정 잠재코드로 매핑을 학습하며, 학습 완료 후 추론단계에서 임의의 음성 신호를 바탕으로 임의의 얼굴 모델에 대해 음성 신호에 동기화된 3차원 얼굴 발화 애니메이션 생성이 가능한 완전한 엔드투엔드(end-to-end) 파이프라인으로 동작한다. 실험을 통해 본 방법이 기존 연구 및 다른 설계 대안보다 우수한 성능을 보였으며, 학습 시 접하지 않은 형태 또는 메쉬 구조를 가진 임의의 얼굴 모델에도 강건하게 작동 가능함을 확인하였다. 마지막으로, 본 방법은 직관적인 표정 편집 및 상부 얼굴 표정을 유지하면서 입 모양만 변경하는 메쉬 더빙 등의 실용적인 응용이 가능하다.

본 방법은 FACS기반의 잠재코드를 활용하기 때문에 얼굴 움직임을 세분화할 뿐만 아니라 이를 조합하여 행복, 슬픔, 공포 등과 같은 다양한 감정 표현을 구성할 수 있다는 장점이 있다. 그러나, 이를 효과적으로 활용하기 위해서는 사용자가 FACS에 대한 이해와 이를 기반으로 한 표정 생성 과정에 대한 학습이 필요하다. 이에 향후 연구를 텍스트 입력을 기반으로 FACS 기반의 해석 가능한 잠재 코드를 선택적으로 조합하여 다양한 감정을 표현하는 방향으로 확장하고자 한다. 특히, 텍스트와 잠재 코드 간의 매핑을 통해 텍스트 입력만으로 직관적인 표정 편집이 가능하도록 한다면 차후 다양한 분야에서의 응용이 가능할 것으로 기대된다.

9 감사의 글

본 연구는 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행되었음. (No.RS-2024-00439499, 다양한 속도의 발화 및 맥락 기반 감정 표현이 가능한, 극사실적부터 고도로 스타일화된 얼굴 아바타 생성 기술)

References

[1] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," vol. 36, no. 4, pp. 1–12, 2017.

[2] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," 2019.

[3] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "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.

[4] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "Faceformer: Speech-driven 3d facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 770–18 780.

[5] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T.-T. Wong, "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. 12 780–12 790.

[6] B. Thambiraja, I. Habibie, S. Aliakbarian, D. Cosker, C. Theobalt, and J. Thies, "Imitator: Personalized speech-driven 3d facial animation," 2023.

[7] D. Qin, J. Saito, N. Aigerman, G. Thibault, and T. Komura, "Neural face rigging for animating and retargeting facial meshes in the wild," in *SIGGRAPH 2023 Conference Papers*, 2023.

[8] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[9] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[10] R. Li, K. Bladin, Y. Zhao, C. Chinara, O. Ingraham, P. Xiang, X. Ren, P. Prasad, B. Kishore, J. Xing, and H. Li, "Learning formation of physically-based face attributes," 2020.

[11] C.-h. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, E. Brockmeyer, L. Evans, T. Godisart, H. Ha, X. Huang, A. Hypes, T. Koska, S. Krenn, S. Lombardi, X. Luo,

K. McPhail, L. Millerschoen, M. Perdoch, M. Pitts, A. Richard, J. Saragih, J. Saragih, T. Shiratori, T. Simon, M. Stewart, A. Trimble, X. Weng, D. Whitewolf, C. Wu, S.-I. Yu, and Y. Sheikh, "Multiface: A dataset for neural face rendering," in *arXiv*, 2022.

[12] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. Black, "Capture, learning, and synthesis of 3D speaking styles," 2019, pp. 10 101–10 111.

[13] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3d face analysis," vol. 101, no. 3, pp. 437–458, February 2013.

[14] M. Brand, "Voice puppetry," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 21–28.

[15] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," vol. 37, no. 4, pp. 1–10, 2018.

[16] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: an animator-centric viseme model for expressive lip synchronization," *ACM Transactions on graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[17] M. Villanueva Aylagas, H. Anadon Leon, M. Teye, and K. Tollmar, "Voice2face: Audio-driven facial and tongue rig animations with cvaes," in *Computer Graphics Forum*, vol. 41, no. 8. Wiley Online Library, 2022, pp. 255–265.

[18] S. Medina, D. Tome, C. Stoll, M. Tiede, K. Munhall, A. G. Hauptmann, and I. Matthews, "Speech driven tongue animation," June 2022, pp. 20 406–20 416.

[19] J.-y. Noh and U. Neumann, "Expression cloning," in *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001, pp. 277–288.

[20] R. W. Sumner and J. Popović, "Deformation transfer for triangle meshes," *ACM Transactions on graphics (TOG)*, vol. 23, no. 3, pp. 399–405, 2004.

[21] Q. Tan, L. Gao, Y.-K. Lai, and S. Xia, "Variational autoencoders for deforming 3d mesh models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5841–5850.

[22] L. Gao, J. Yang, Y.-L. Qiao, Y.-K. Lai, P. L. Rosin, W. Xu, and S. Xia, "Automatic unpaired shape deformation transfer," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 6, pp. 1–15, 2018.

[23] S. Kim, S. Jung, K. Seo, R. B. i Ribera, and J. Noh, "Deep learning-based unsupervised human facial retargeting," in *Computer Graphics Forum*, vol. 40, no. 7. Wiley Online Library, 2021, pp. 45–55.

[24] L. Moser, C. Chien, M. Williams, J. Serra, D. Hendler, and D. Roble, "Semi-supervised video-driven facial animation transfer for production," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 6, pp. 1–18, 2021.

[25] N. Aigerman, K. Gupta, V. G. Kim, S. Chaudhuri, J. Saito, and T. Groueix, "Neural jacobian fields: learning intrinsic mappings of arbitrary meshes," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–17, 2022.

[26] "Live link face," https://apps.apple.com/us/app/live-link-face/id1495370836.

[27] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 914–921.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[29] C. Loop, "Smooth subdivision surfaces based on triangles," 1987.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, 2017.

[31] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," vol. 30, 2017.

[32] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[33] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *ECCV*, 2020.

[34] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[35] S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., *Computer Vision – ECCV 2022*. Springer, 2022.

[36] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3d faces using convolutional mesh autoencoders," September 2018.

[37] N. Sharp, S. Attaiki, K. Crane, and M. Ovsjanikov, "Diffusionnet: Discretization agnostic learning on surfaces," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 3, pp. 1–16, 2022.

〈저자소개〉

서광균

- 2011 년~2016 년: KAIST 전기및전자공학부 학사
- 2016 년~2018 년: KAIST 문화기술대학원 석사
- 2018 년~2024 년: KAIST 문화기술대학원 박사
- 2024 년~현재: Flawless AI (Research Scientist)
- 관심분야: Facial Animation, Generative Models
- https://orcid.org/0000-0003-0570-4915

차시헌

- 2013년~2020년: 한국예술종합학교 미술원 조형예술과 예술사
- 2022년~2022년: KAIST 문화기술대학원 석사
- 2022년~현재: KAIST 문화기술대학원 박사과정
- 관심분야: Computer Graphics, Facial Animation, Speech Animation, Facial Modeling
- https://orcid.org/0000-0001-9506-9438

나현호

- 2016년~2021년: 고려대학교 전기전자공학부 학사
- 2021년~2023년: KAIST 문화기술대학원 석사
- 2023년~현재: KAIST 문화기술대학원 박사과정
- 관심분야: Computer Graphics, Facial Animation, Speech Animation
- https://orcid.org/0000-0002-6818-1595

이인엽

- 2021년~2023년: 고려대학교 컴퓨터학과 학사
- 2023년~현재: KAIST 문화기술대학원 석사과정
- 관심분야: Computer Graphics, Facial Animation, Animation Editing, Facial Reconstruction
- https://orcid.org/0009-0003-6014-6132

노준용

- 1990년~1994년 University of Southern California Electrical Engineering 학사
- 1994년~1996년 University of Southern California Computer Engineering 석사
- 1996년~2002년 University of Southern California Computer Science 박사
- 2003년~2006년 Rhythm and Hues Studio, Graphics Scientist
- 2006년~현재 카이스트 문화기술대학원 교수
- 2011년~현재 카이스트 석좌 교수
- 2012년~현재 카이스트 전산학과 겸임교수
- 2016년~2020년 카이스트 문화기술연구소 소장
- 2016년~2020년 카이스트 문화기술대학원 학과장
- 관심분야: Computer Graphics, Computer Vision, Facial Modeling, Facial Animation, Character Animation, Image & Video Manipulation/Generation
- https://orcid.org/0000-0003-1925-3326