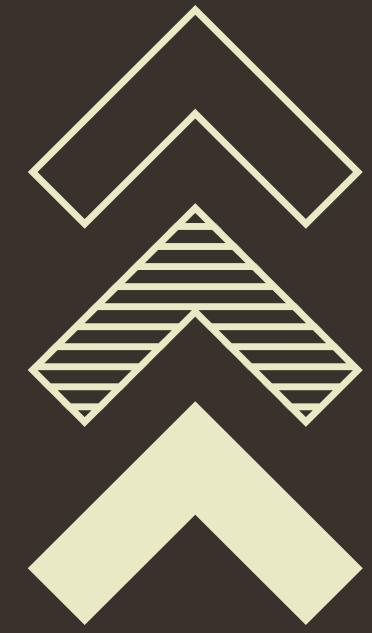


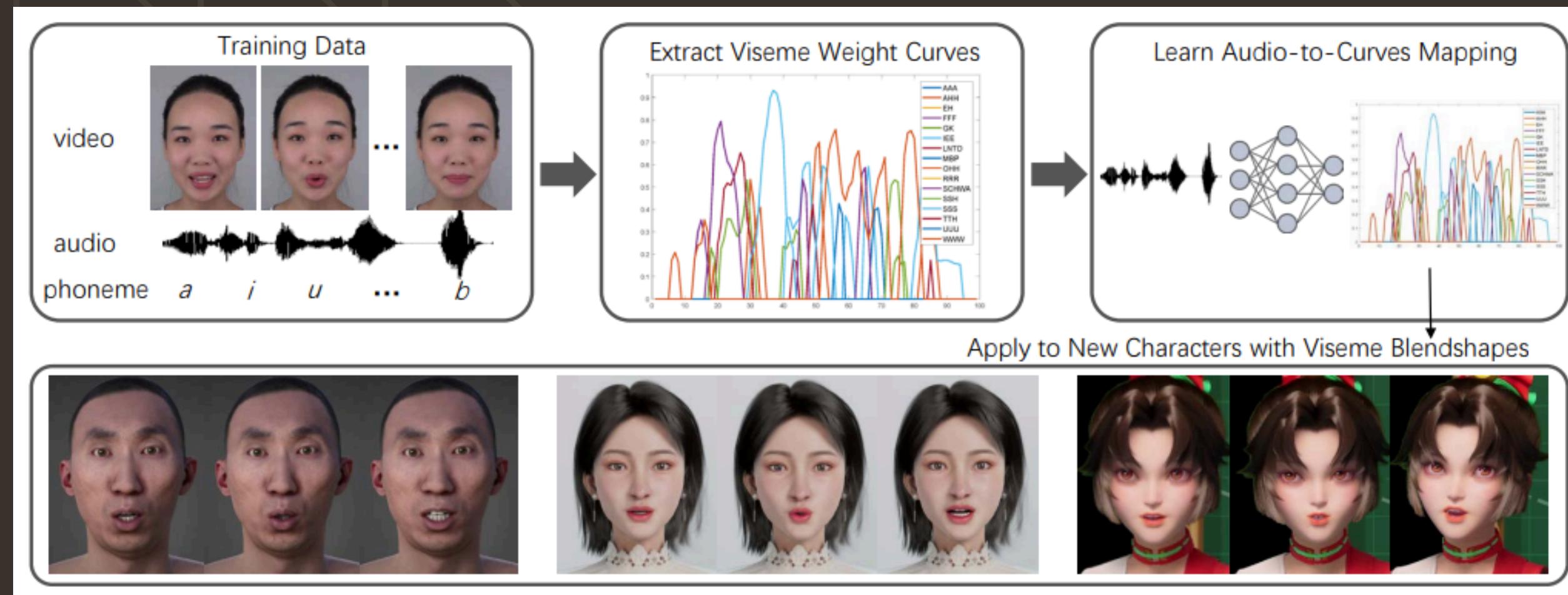
Sogang University

LEARNING AUDIO-DRIVEN VISEME DYNAMICS FOR 3D FACE ANIMATION



이혜민
서강대학교 메타버스전문대학원

FACIAL ANIMATION TECHNIQUE



PHONEME-GUIDED 3D FACIAL TRACKING: 음소 정보를 활용해 얼굴의 동작을 추적

LEARNING AUDIO-TO-CURVES MAPPING: 오디오 입력에서 직접 Viseme 곡선을 예측

SPEECH ANIMATION PRODUCTION: Viseme 곡선을 바탕으로 Blendshape 획득하여
고품질의 애니메이션을 생산

INTRODUCTION

Vertex-Based Animation



[Karras et al. 2017]



VOCA
[Cudeiro et al. 2019]



MeshTalk
[Richard et al. 2021]



FaceFormer
[Fan et al. 2022]

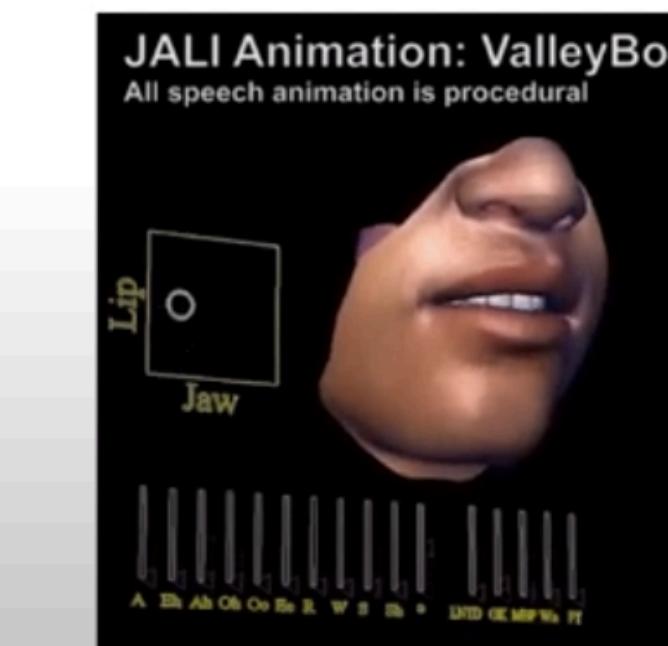
Parameter-Based Animation



[Taylor et al. 2017]



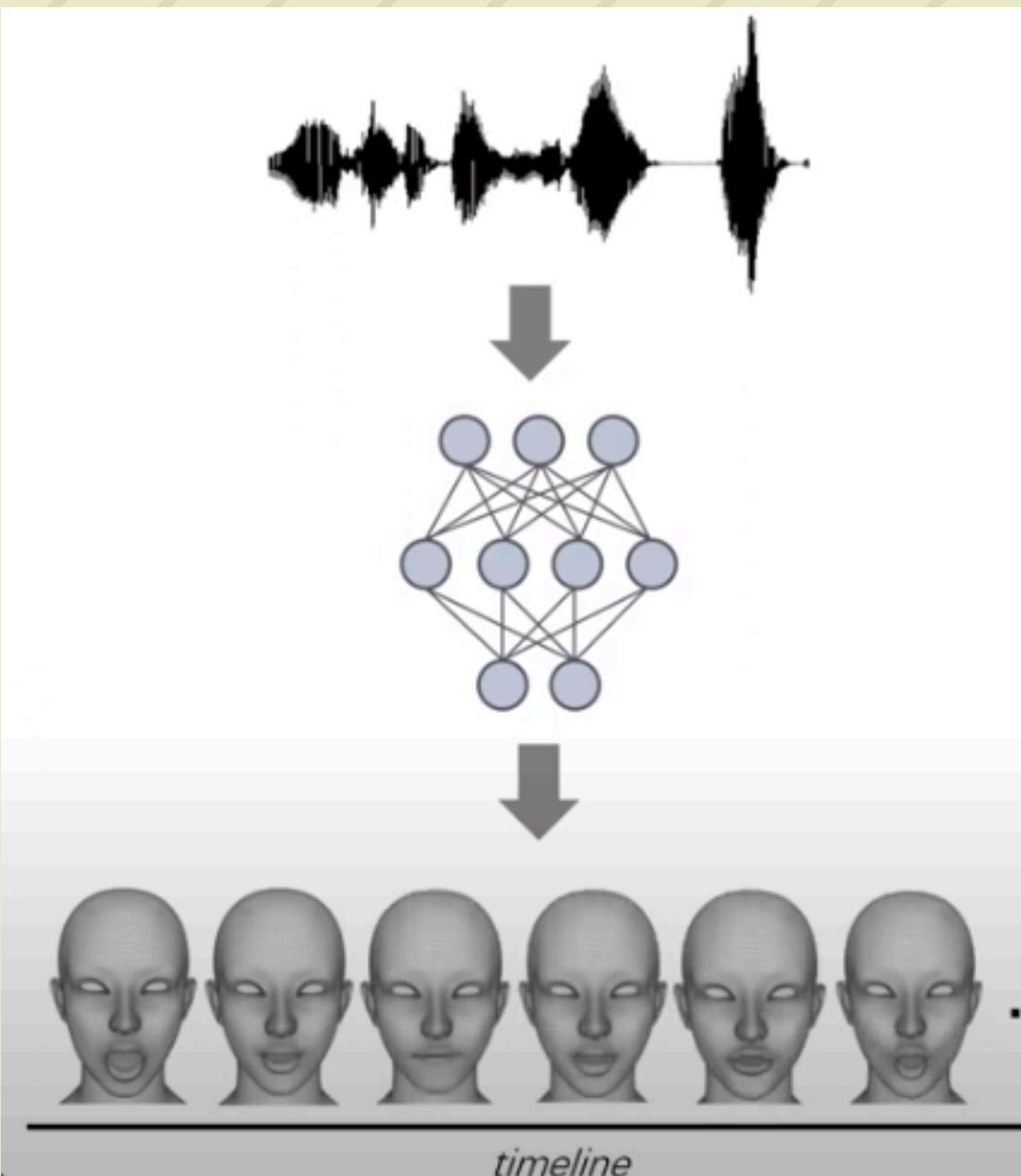
VisemeNet
[Zhou et al. 2018]



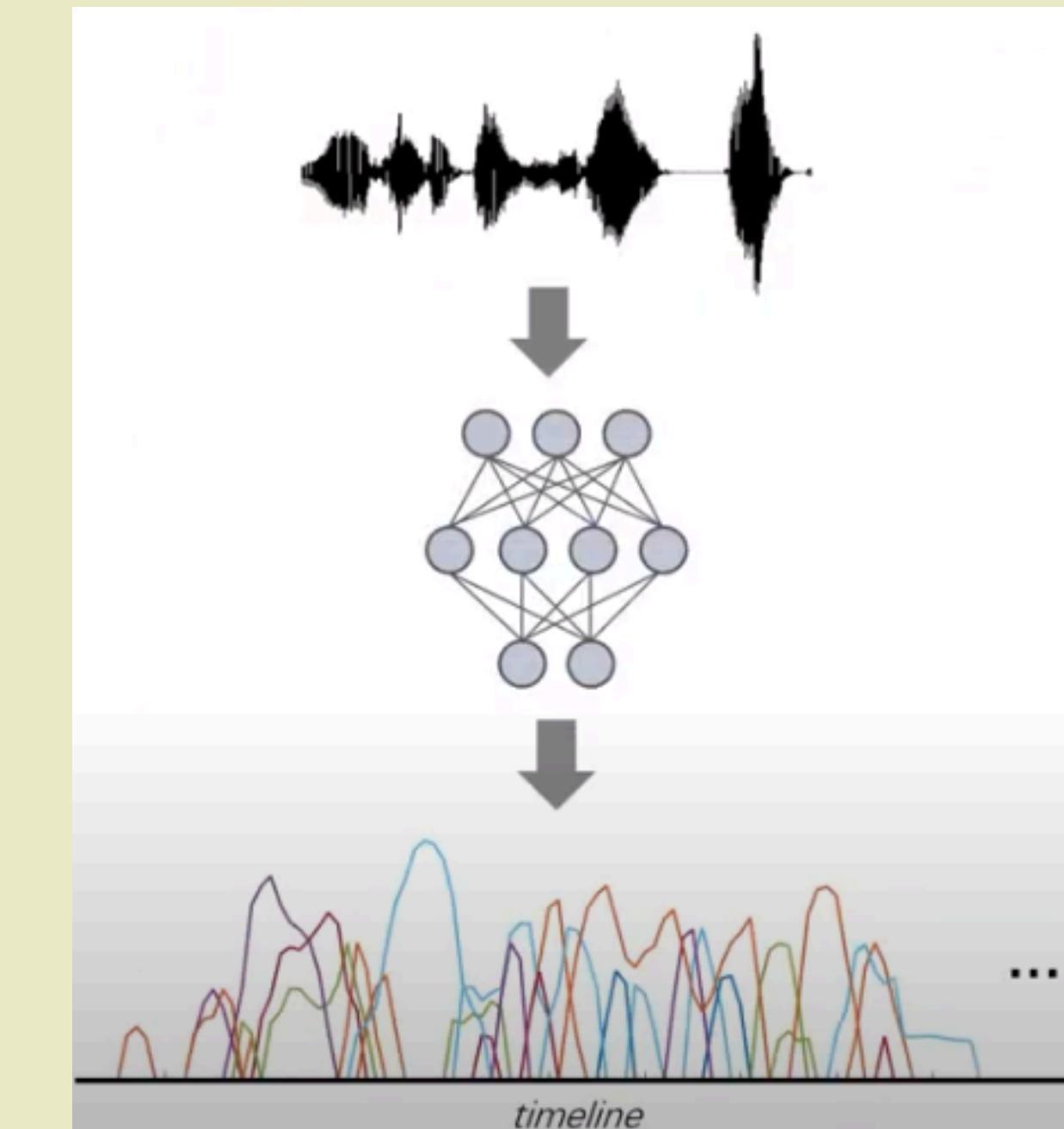
JALI [Edwards et al. 2016]

INTRODUCTION

Vertex-Based Animation



Parameter-Based Animation





Facial Performance Capture from Video

Blendshape과 multi-linear 모델을 통해 비디오의 프레임 분석, 얼굴 표정 구현 학습 기반 방법을 개발하여 얼굴 표정을 더 빠르고 정확하게 파악함으로써 효율성 증대

Parameter-Based 3D Facial Animation from Audio

Independent Component Analysis을 이용해 음소와 얼굴 움직임을 매핑, 보다 자연스러운 얼굴 표현을 위해 딥러닝 기법을 적용

Vertex-Based 3D Facial Animation from Audio

딥러닝 학습을 기반으로 오디오로부터 얼굴 모델의 버텍스 위치에 대한 매핑을 직접 생성함
VOCA와 MeshTalk가 있지만 예술가 친화적이지 않고 애니메이션 제작 워크플로우에 통합 어려움

Face Vid eo D at a se

Dataset

중국인 여배우가 말하는 16시간 동영상 (12,000개의 발화 중 10,000개는 훈련용)
대부분 중국어로 이루어져 있지만 일부는 영어 단어도 포함됨
음소는 사내 도구를 사용하여 시작과 종료 타임스탬프와 함께 생성됨
[McAuliffe et al. 2017].

Viseme	Phoneme (English)	Phoneme (Chinese)	Viseme	Phoneme (English)	Phoneme (Chinese)
AAA	æ, eɪ, aɪ	ai	OHH	əʊ, ɔ:, ɔɪ	o
AHH	ɒ, a:, ʌ, aʊ	a	RRR	r	r
EH	ɛ, h	ei	SCHWA	ə, ɜ:	e, er
FFF	f, v	f	SSH	tʃ, ʃ, ʒ, dʒ	zh, ch, sh
GK	g, k	g, k, h	SSS	ð, s, z	z, c, s
IEE	i, i:, j	i	TTH	θ	j, q, x
LNTD	l, n, t, d	l, n, t, d	UUU	ʊ, u:	iu
MBP	m, b, p	m, b, p	WWW	w	u, v

Visemes & Phonemes

16개의 Viseme과 새로운 음소 가이드 얼굴 추적 알고리즘을 이용해
viseme의 가중치 계산 후 오디오-투-커브 신경망 모델을 훈련
그리고 새 캐릭터에 오디오 구동 애니메이션을 제작함

PHONEME-GUIDED 3D FACIAL TRACKING

Procedural Viseme Weights Generation

각 오디오 트랙의 프레임 별로
음소의 발음 시간과 유형에 따른
Viseme 가중치 생성

:JALI 알고리즘과 유사한 방식

JALI Animation: ValleyBoy
All speech animation is procedural

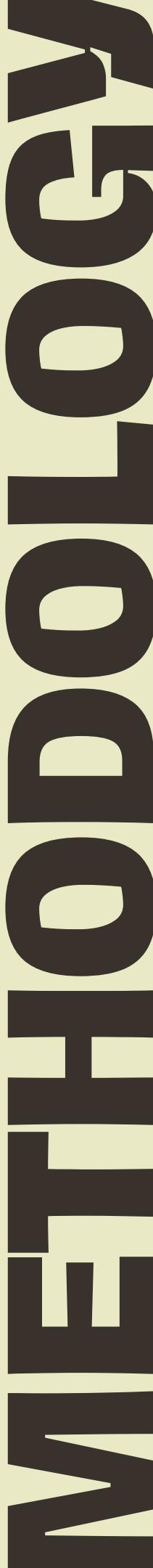


Viseme Blendshape Preparation

배우의 중립 3D 얼굴 모델과 16개의 Viseme BlendShape 구성

: 배우의 얼굴에서 중립 상태의 프레임 추출 후 3D 얼굴 모델링 기술을 이용해 3D 모델과 텍스처 생성

Viseme 템플릿 모델을 3D 모델에 맞게 조정 및 최적화



Phoneme-Guided Viseme Parametric Fitting

$$L(\mathcal{X}^j) = w_1 L_{lmk} + w_2 L_{rgb} + w_3 L_{sup} + w_4 L_{act} \\ + w_5 L_{flow} + w_6 L_{diff} + w_7 L_{range},$$

3DMM이라는 3D 모델링 알고리즘에 근거한 것으로,
다양한 손실 함수를 통해 정밀한 얼굴 애니메이션을 생성함
손실 함수는 모델의 예측이 실제 데이터와 얼마나 잘 일치 하는지를 측정하는 함수
 $w1 = 0.8, w2 = 1.0, w3 = 800, w4 = 150, w5 = 1.0, w6 = 300, w7 = 100$ 으로 설정

Landmark Loss(Llmk): 특정 표식(landmark) 위치의 정확성을 측정

RGB Photo Loss(Lrgb) : 색상 정보의 정확성을 측정

Phoneme-Guided Suppression Loss(Lsup) : 특정 프레임에서 활성화되지 않게 설계

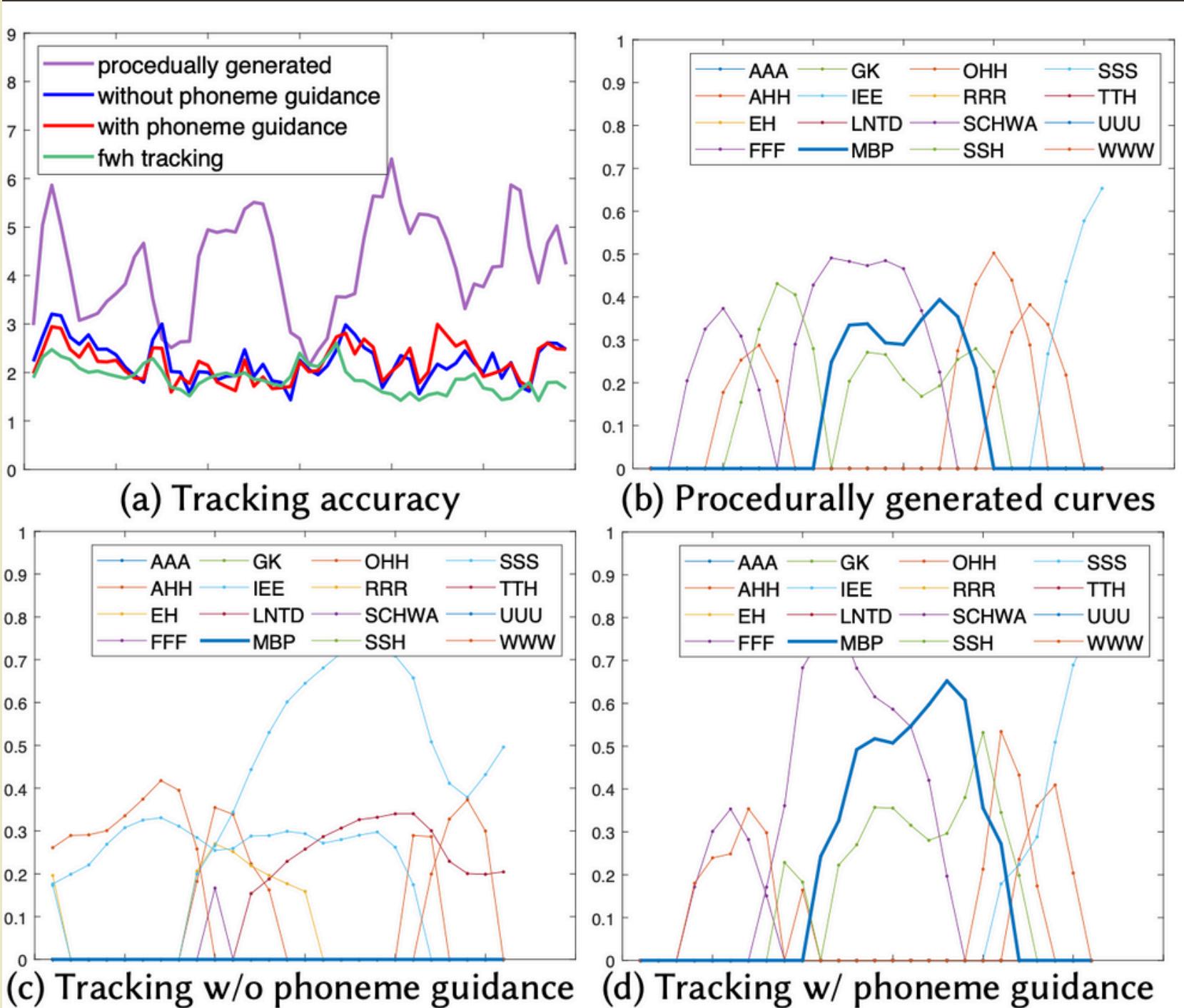
Activation Loss(Lact): 필요한 Viseme를 적절히 활성화하기 위해 사용

Optical Flow Loss(Lflow): 연속된 두 프레임 간의 이미지 변화를 추적

Temporal Consistency Loss(Ldiff): 시간에 따른 애니메이션의 일관성을 보장

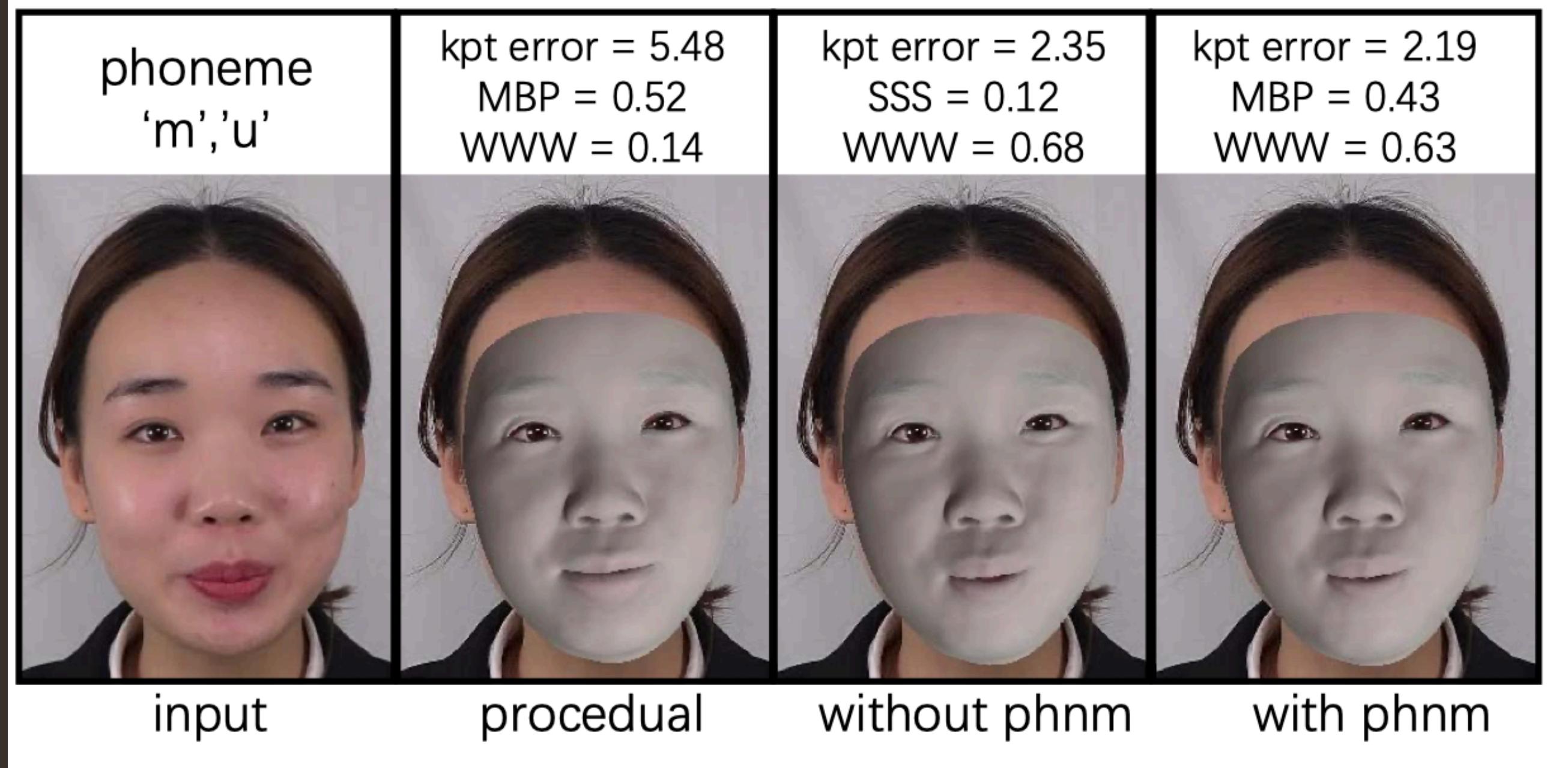
Value Range Loss(Lrange): 모델의 출력 값이 특정 범위 내에 있도록 제한

Figure 3

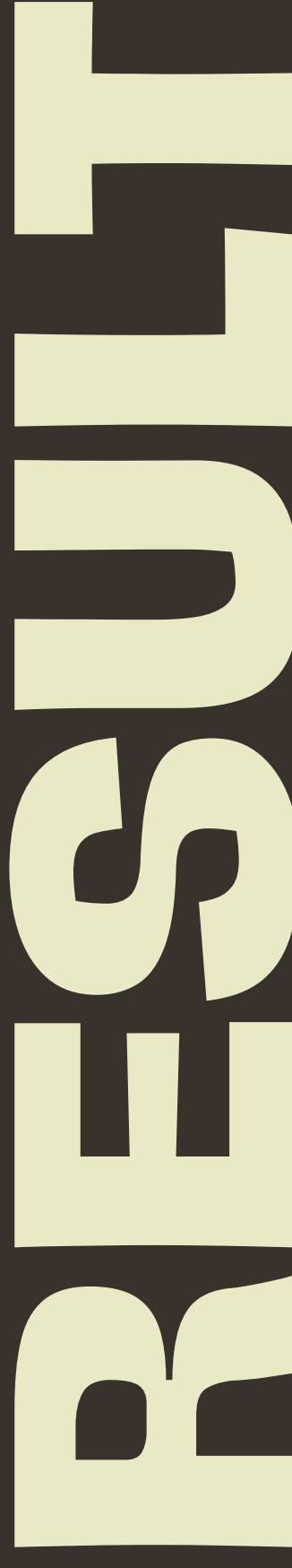


1. 추적 정확성 평가: 입 주변의 주요 특징점들의 오차를 평균적으로 보여줌
2. FaceWarehouse와의 정확성 비교: FaceWarehouse를 사용하는 기존의 알고리즘과 비슷한 수준의 정확도를 보여줌
3. 음소 지도의 영향: 음소 지도는 추적 알고리즘의 정확성 자체에는 직접적인 영향을 미치지 않지만, 실제로는 훨씬 정확하게 움직임을 추적 가능
4. 자연스러운 움직임 재현: 연구팀이 추적한 비선곡선은 일반화되거나 표준화된 형태보다 더 세밀하고 자세한 변화를 포함

Figure 4



LEARNING AUDIO-TO-CURVES MAPPING



Audio Encoder

Wav2Vec2 네트워크 구조 $\{\phi, \varphi\}$

ϕ : 7개의 TCN layer로 구성되며, 오디오 데이터 U_0 를 초기 특징으로 변환

φ : 24개의 멀티헤드 셀프 어텐션 레이어로 구성되며, ϕ 에서 변환된 특징을 문맥화된 음성 표현 U_1 으로 변환

오디오 신호에서 유의미한 정보를 추출하고, 고차원에서 문맥적으로 풍부한 형태로 인코딩

선형 및 보간 레이어

선형 레이어 f_{proj} : U_1 의 특징 차원을 1024에서 512로 감소

보간 레이어 f_{interp} : 특징의 프레임 속도를 viseme 곡선과 일치시키기 위해 사용

이 단계들은 모델이 처리하기 쉽도록 데이터의 차원을 줄이고, 다음 단계의 네트워크 입력으로 적합하게 만듦

주기적 위치 인코딩 (PPE)

$U_3 = U_2 + P$: 오디오 특징 U_2 에 주기적 위치 인코딩을 추가하여 시간적 문맥을 강화
위치 인코딩은 모델이 시퀀스 데이터에서 시간적 위치를 더 잘 이해하도록 유도

LEARNING AUDIO-TO-CURVES MAPPING

Viseme Curves Decoder

Viseme 가중치 매핑

$$Y = fFC(fBLSTM(U3))$$

양방향 LSTM을 거친 후, 완전연결 레이어를 통해 최종적으로 16차원의 viseme 가중치 Y 예측
오디오 특징을 기반으로 정확한 입 모양의 동작을 생성하는 데 필요한 출력을 제공

L1 손실 함수

추적된 viseme 가중치 X 와 예측된 가중치 Y 사이의 L1 손실을 사용하여 모델을 훈련.
L1 손실은 예측과 실제 값 사이의 절대 차이를 최소화하는 데 도움을 줌

Training

SpecAugment

데이터 증강을 통해 모델의 일반화 능력을 향상

Adam 최적화

고정된 학습률로 모델을 최적화

Results and Evaluation

Robustness Evaluation: 시스템이나 모델이 다양한 환경 조건이나 입력 값의 변동성에 대해 얼마나 견딜 수 있는지 평가하는 과정

1. Validation (검증):

- 정의: 이 지표는 훈련된 모델이 검증 데이터셋에서 수행하는 정도
- 계산: 모델의 출력과 검증 데이터셋의 실제 값 사이의 오차(주로 L1 오차)를 계산

2. Volume (볼륨):

- 정의: 오디오 데이터의 볼륨을 조정하여 모델의 볼륨 변화에 대한 강건성을 평가
- 계산: -10 dB에서 +10 dB 사이에서 임의로 조정한 후, 모델을 평가하고 오차를 계산

3. Pitch (음높이):

- 정의: 오디오 데이터의 음높이를 변경하여 음높이 변화에 얼마나 강건한지를 평가
- 계산: 원본 오디오의 음높이를 -3 세미톤에서 +3 세미톤 사이에서 변경한 후, 모델을 평가하고 오차를 계산

4. Speed (속도):

- 정의: 오디오 데이터의 재생 속도를 변경하여 속도 변화에 대해 얼마나 잘 대응하는지를 평가
- 계산: 오디오 데이터의 속도를 0.7배에서 1.3배 사이에서 조정한 후, 모델을 평가하고 오차를 계산

5. Noise (소음):

- 정의: 오디오 데이터에 잡음을 추가하여 모델이 소음이 있는 환경에서 얼마나 잘 작동하는지를 평가
- 계산: 오디오 데이터에 가우시안 잡음(진폭이 0.005 이하)을 추가한 후, 모델을 평가하고 오차를 계산

Table 1. Ablation study for different audio features. The listed values are L1 reconstruction errors multiplied by 10^2 . *Volume* stands for the generality test performed on the validation set with augmented volume of the test audio. The same goes for *pitch*, *speed* and *noise*.

Feature \ Metric	validation	volume	pitch	speed	noise
FBank	3.21	3.32	4.59	3.39	4.42
LPC	5.26	5.27	6.50	5.35	5.93
PPG [Huang et al. 2021]	3.01	3.01	3.07	7.50	3.13
Wav2Vec2 ¹ [Baevski et al. 2020]	1.88	1.89	2.48	2.29	2.47
Wav2Vec2 ² [Baevski et al. 2020]	1.82	1.83	2.55	2.19	2.27
Wav2Vec2 ³ (ours) [Conneau et al. 2020]	1.77	1.77	2.44	2.06	2.09

Table 2. Ablation study for different decoder backbones. The listed values are L1 reconstruction errors multiplied by 10^2 .

Decoder \ Metric	validation	volume	pitch	speed	noise
TCN	1.77	1.78	2.51	2.11	2.11
LSTM	1.77	1.78	2.50	2.12	2.16
Transformer	1.79	1.80	2.54	2.13	2.14
BLSTM (ours)	1.77	1.77	2.44	2.06	2.09

Ablation Study on Audio Features

각 오디오 특징의 복원 오류를 L1 측정 방법으로 계산한 값($\times 10^2$)을 보여줌

이 오류 값은 소리의 다양한 속성(볼륨, 음높이, 속도, 소음)에 따라 어떻게 변화하는지를 나타냄
Wav2Vec23 모델이 다른 모델들에 비해 일관되게 낮은 오류 값을 보여주고 있어, 다양한 언어를 포함한 데이터셋에서의 사전 훈련이 효과적

Ablation Study on Decoder Backbones

TCN, LSTM, Transformer, BLSTM들이 오디오 특징을 시각적 표현으로 변환하는 데 얼마나 효과적인지를 비교

BLSTM이 이웃 정보를 효과적으로 집계하므로 우수한 성능 보임

**Figure 5 &
Table 3**

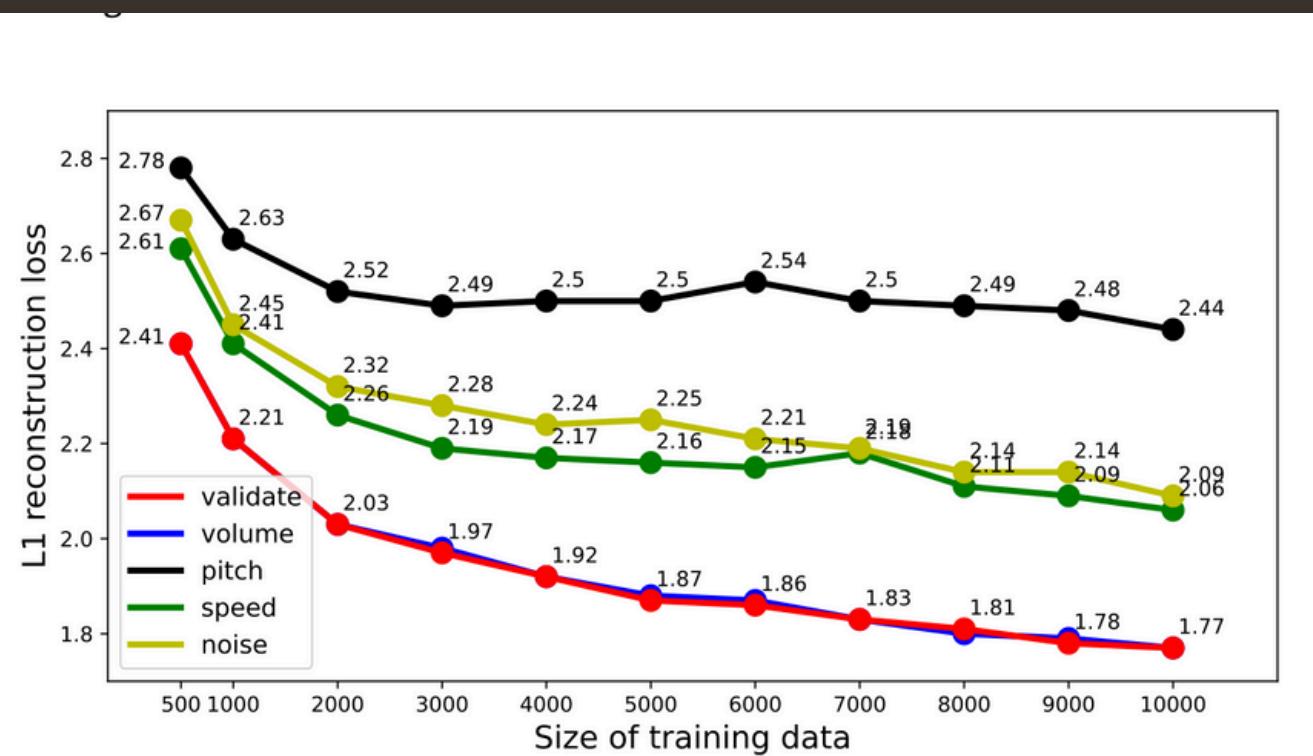


Fig. 5. L1 reconstruction losses on validation set for our model trained on datasets with various sizes. All experiments are trained for 10^6 iterations.

Table 3. Comparison with state-of-the-art deep models. The listed values are L1 reconstruction errors multiplied by 10^2 .

Method \ Metric	validation	volume	pitch	speed	noise
Karras et al. [2017]	4.11	4.15	5.79	7.42	6.34
Huang et al. [2021]	2.38	2.38	2.45	7.43	2.69
FaceFormer [Fan et al. 2022]	2.10	2.10	2.69	2.51	2.89
Ours	1.77	1.77	2.44	2.06	2.09

Size of Training Dataset

전체 데이터로 훈련시킨 우리 모델이 가장 좋은 성능을 보여주지만, 예를 들어 3000 발화로 훈련된 모델도 합리적으로 좋은 결과를 제공

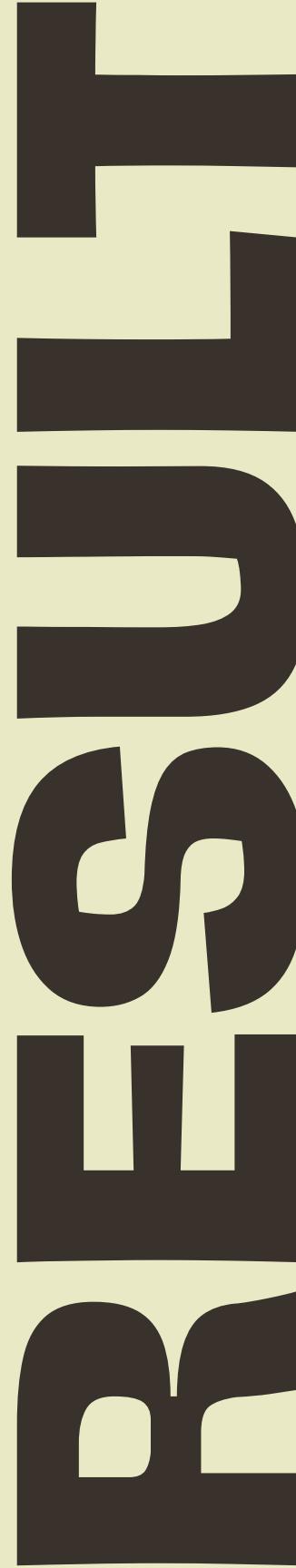
Comparison with State-of-the-arts

다양한 오디오 변화(음높이, 속도 등)를 처리할 때 최신 오디오 기반 음성 애니메이션 모델과 비교하여 우리 모델이 가장 좋은 성능을 달성

Generalize to Unseen Speakers and Other Languages

강력한 사전 훈련된 오디오 특성 모델 덕분에 훈련 데이터와 다른 화자나 외국어에도 적용 가능

SPEECH ANIMATION PRODUCTION



Viseme Blendshape Scanning

배우가 다양한 음소를 발음하는 모습을 간단한 카메라 어레이 시스템을 사용해 멀티뷰 비디오를 캡처

Beeler et al.[2010]과 비슷한 방식으로 텍스처 및 노멀 맵 추출

애니메이션의 동작 정확성 검증

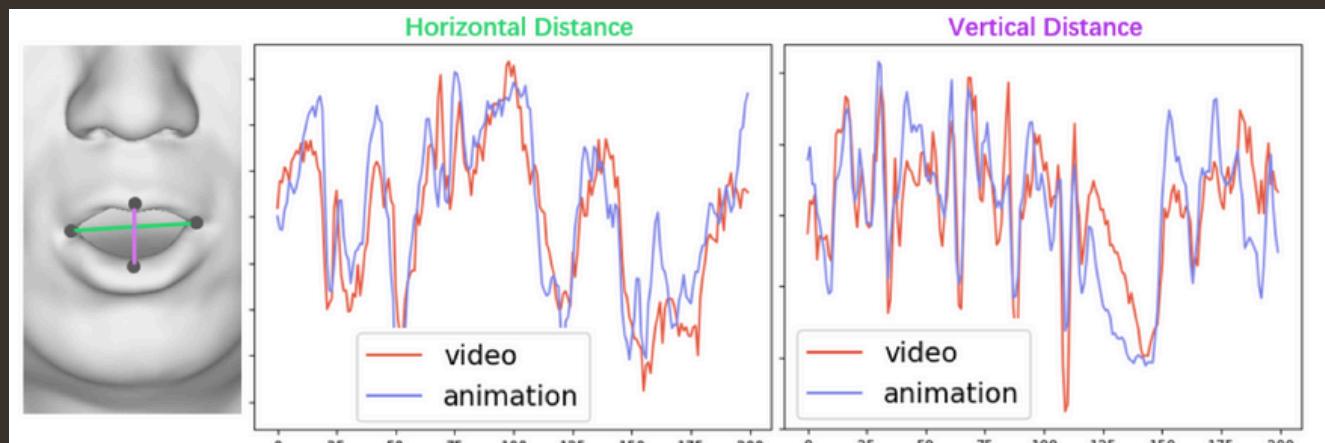


Fig. 8. Comparison of lip motion curves between real video and animation. We record the temporal sequences of the horizontal and vertical distances between two pairs of keypoints from both real video and animation. The comparison demonstrates that the lip motions in the produced animation well resemble real lip motions.

표정과 함께하는 음성 애니메이션 생성

Retargeting and Bone Animation

Retargeting: 이미 존재하는 애니메이션 데이터나 형태를 다른 캐릭터나 모델에 적용할 때 사용되는 기술로 논문에서는 데포메이션 전달 알고리즘(Deformation Transfer Algorithm)을 사용해 유연성 및 최적화 추구

하지만 많은 리소스에 민감한 응용 프로그램에서는 Bone Animation이 선호

Bone Animation: 캐릭터의 뼈대를 사용하여 애니메이션을 제어하는 기술로 논문에서는 SSDR 알고리즘 [Le and Deng 2012]을 사용해 효율성 추구

3D voi lcul tive syn thesis



Conclusion

오디오로부터 3D 인물의 사실적인 립싱크가 포함된 발화 애니메이션 구현
viseme 기반 매개변수 공간 애니메이션 곡선 제안
-> 예술가 친화적이면서, 새롭고, 다양한 스타일의캐릭터에 활용 가능함
게임 업계, AI 기반 디지털 휴먼 애플리케이션에 적용 가능

Limitation

발화 시 혀의 움직임은 표착 어려움
viseme blendshape에 혀와 치아의 정적 위치를 연결, viseme 곡선을 통
해 애니메이션 구현
입안을 밝게 할 시 부자연스러움



THANK
YOU

