Facial Animation :
Audio-driven 3D Lip-sync Animation

20240820 Facial Animation team 박혜원

Schedule

- 1주차
  - 프로젝트 세팅 및 온보딩
- 2주차~14주차
  - 머신 러닝 / 딥러닝 학습
  - Viseme Model 진행
- 15~23주차
  - Audio2Blendshape 진행
- 24주차
  - 최종 발표

Contents

Part.1 Viseme Model

01. Introduction

Introduction

* Viseme Lip-sync Model은 **규칙 기반의 동시조음 모델**로, **한국어에 특화**된 자연스러운 립싱크 애니메이션을 생성

* Viseme Lip-sync Model은 Phoneme에서 Viseme으로의 규칙 기반 Mapping을 통해 **다양한 음성으로** 부터 **일관된 립싱크 애니메이션**을 생성하는 것이 목표

Contribution

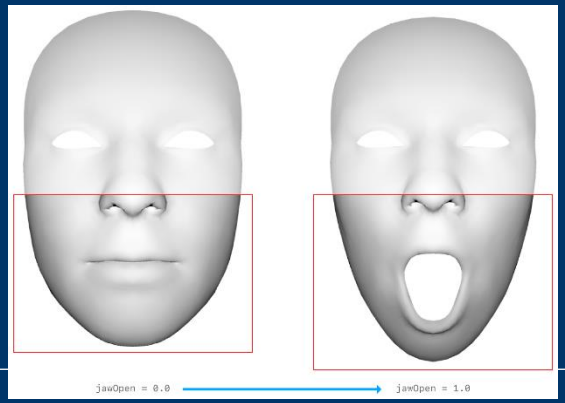
Viseme Lip-sync Model는 PPG를 활용하여 Phoneme 정보를 추출하고, 이를 규칙 기반 Viseme 매핑을 통해 립싱크 애니메이션으로 변환하는 모델

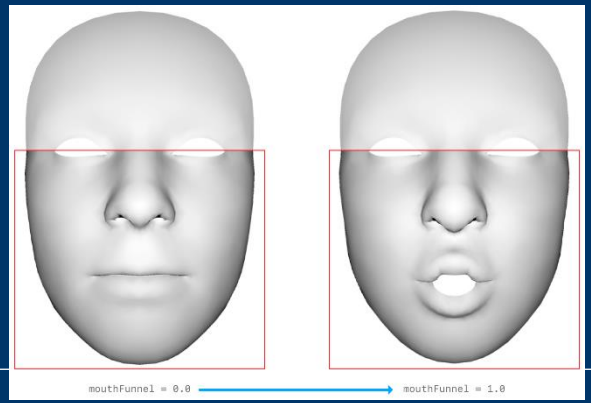
- **Amplitude 정보를 사용해 음성의 다양한 특징을 반영하면서도 정확한 타이밍에 립싱크 애니메이션을 생성할 수 있는 Robust한 모델을 제시**
- **한국어의 언어학적 특성을 반영한 동시조음 모델을 구축 하여 실제 입 모양에 가까운 자연스러운 입 모양 애니메이션을 생성**

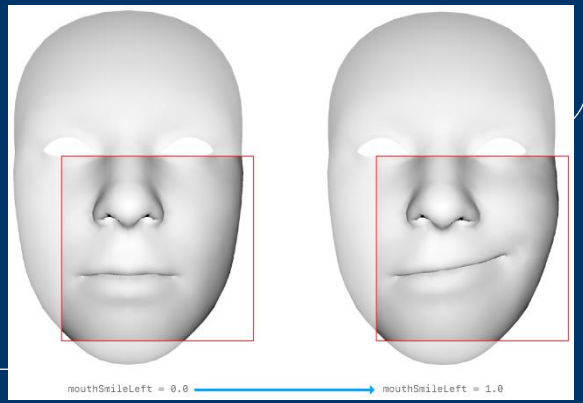
02. Background

ARKit Face Blendshapes

- Blendshape : 기본 얼굴 pose를 blend하여 다양한 표정을 만들어내는 기법
- **ARKit Face Blendshapes** : 얼굴의 주요 근육 움직임과 표정을 나타내는 52개의 Blendshape을 사용하여 복잡한 표정을 구현







Viseme Mapping

- Viseme : 특정 Phoneme에 대응하는 입 모양
- 조음 위치와 조음 방법 등을 통해 Phoneme에서 Viseme 매핑하는 규칙 정의 가능
- ppg가 추출하고 있는 Phoneme (85개)
    - SIL,AA,AA0,AA1,AA2,AE,AE0,AE1,AE2,AH,AH0,AH1,AH2,AO,AO0,AO1,AO2,AW,AW0,AW1,AW2,AY,AY0,AY1,AY2,B,CH,D,DH,EH,EH0,EH1,EH2,ER,ER0,ER1,ER2,EY,EY0,EY1,EY2,F,G,HH,IH,IH0,IH1,IH2,IY,IY0,IY1,IY2,JH,K,L,M,N,NG,OW,OW0,OW1,OW2,OY,OY0,OY1,OY2,P,R,S,SH,T,TH,UH,UH0,UH1,UH2,UW,UW0,UW1,UW2,V,W,Y,Z,ZH
- 사용하고 있는 Viseme (10개)
    - Ah, E, Woo, Eh, Oh, Eu, Eo, PBM1, PBM2, Neutral

Viseme Mapping

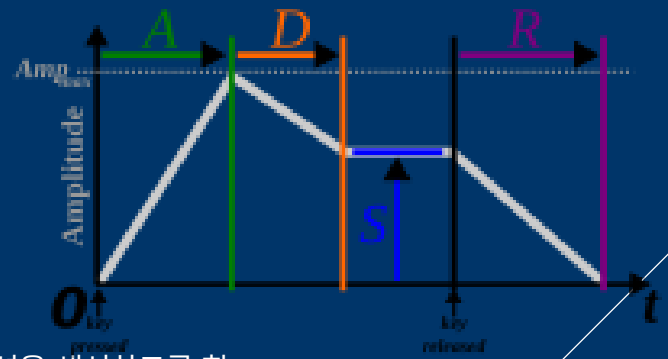
Vowel		Semi Vowel		Stop		Affricate		Fricative		Aspirat		Liquid		Nasal	
AA	ㅏ	W	ㅟ	B	ㅂ	CH	ㅊ	DH	ㄸ	HH	ㅎ	L	ㄹ	M	ㅁ
AE	ㅐ	Y	ㅠ	D	ㄷ	JH	ㅈ	F	ㅍ			R	ㄹ	N	ㄴ
AH	ㅓ	G	ㄱ	G	ㄱ			S	ㅅ					NG	NG
AO	ㅗ	K	ㅋ	K	ㅋ			SH	ㅅ						
AW	ㅏㅜ	P	ㅍ	P	ㅍ			TH	ㄸ						
AY	ㅏㅣ	T	E	T	E			V	ㅂ						
EH	ㅓ							Z	ㅈ						
ER	ㅓ							ZH	ㅈ						
EY	ㅔ														
IH	ㅣ														
IY	ㅣ														
OW	ㅗㅜ														
OY	ㅗㅣ														
UH	ㅠ														
UW	ㅜ														

Viseme Mapping

조음 위치	리스트(EN)
모음(Vowel)	AA,AE,AH,AO,AW,AY,EH,ER,EY,IH,IY,OW,OY,UH,UW,W,Y
양순음(Bilabial)	B,F,M,N,P,V
치조음(Dental)	D,DH,L,R,S,SH,T,TH
경구개음(Palatal)	CH,JH,Z,ZH
연구개음(Velar)	G,K
후음(Glottal)	HH
묵음(SIL)	SIL,NG,SPN

ADSR Envelope

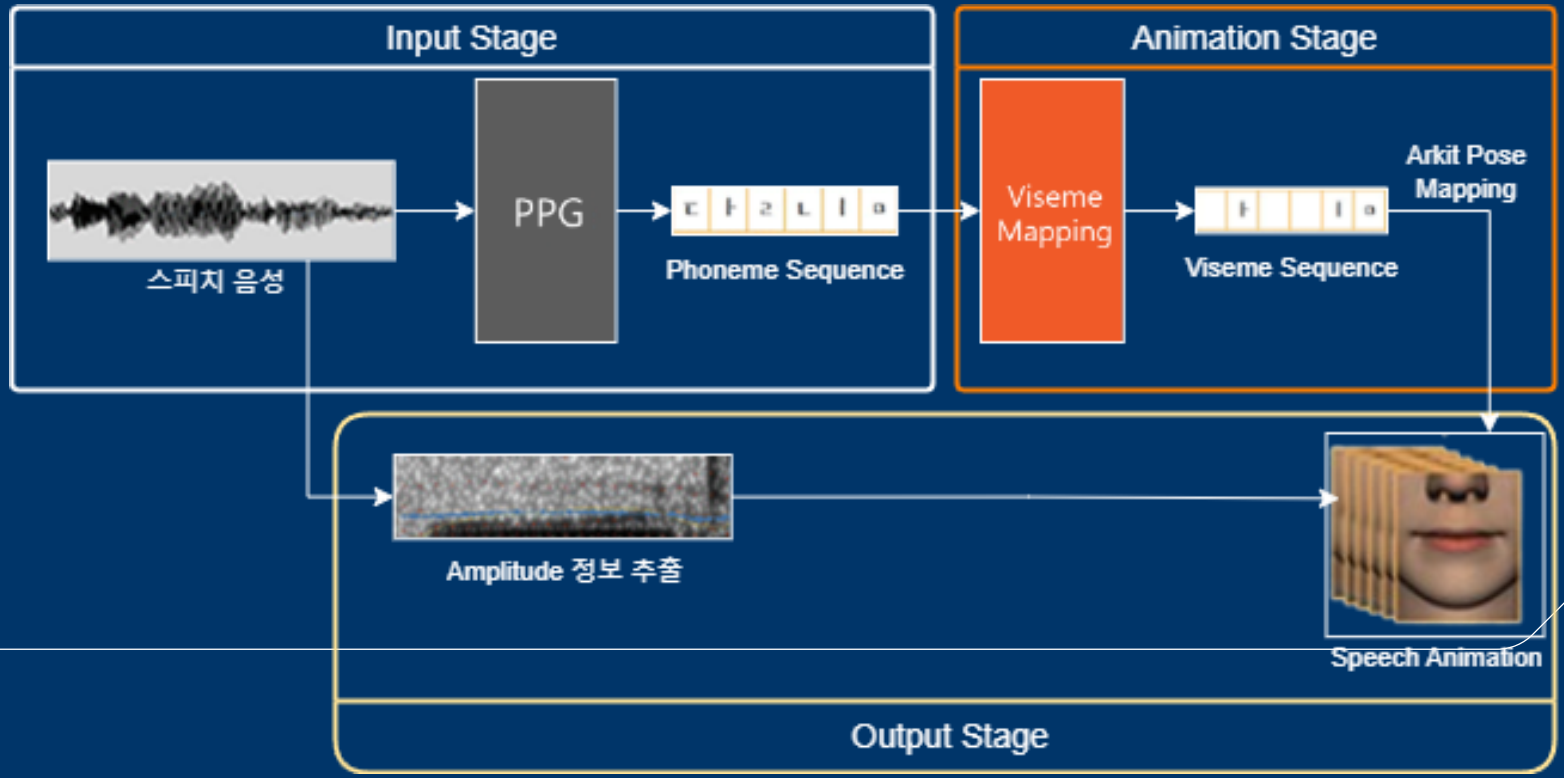
- ADSR(Attack, Decay, Sustain, Release) envelope는 소리의 동적 특성을 제어하는 데 사용되는 주요 요소
- ADSR는 다음 네 가지 단계로 구성
    1. Attack: 소리가 시작되고 최대 레벨에 도달하는 단계
    2. Decay: 공격 단계 후 소리가 Sustain 레벨에 도달하는 단계
    3. Sustain: 소리가 유지되는 단계
    4. Release: 소리가 완전히 사라지는 단계



Viseme Lip-sync Model은 amplitude가 ADSR Curve를 따라 자연스러운 립싱크 애니메이션을 생성하도록 함

03. Model

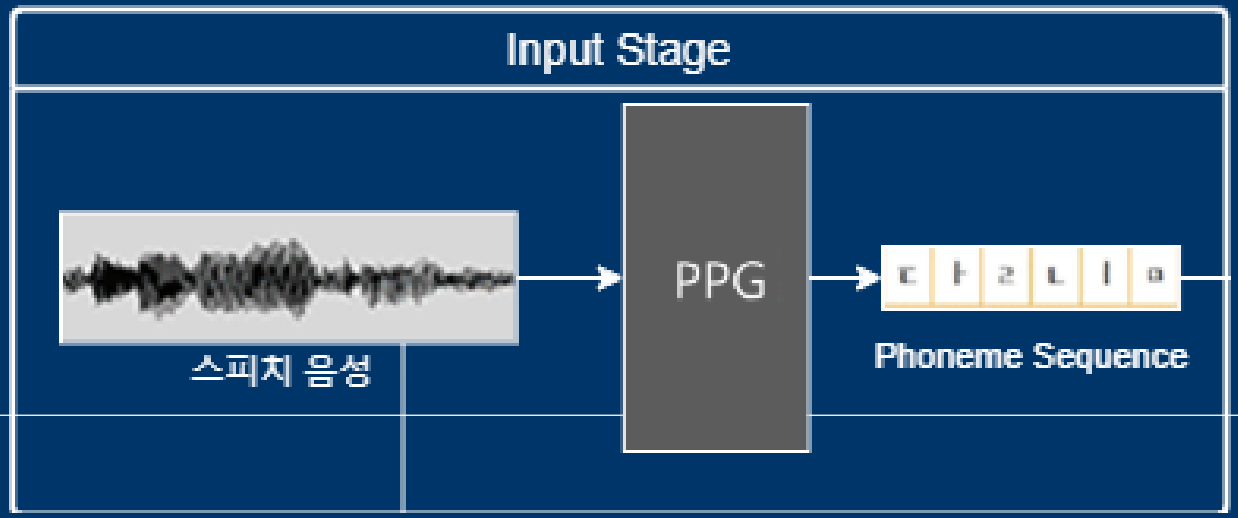
Model structure



Input Stage

PPG

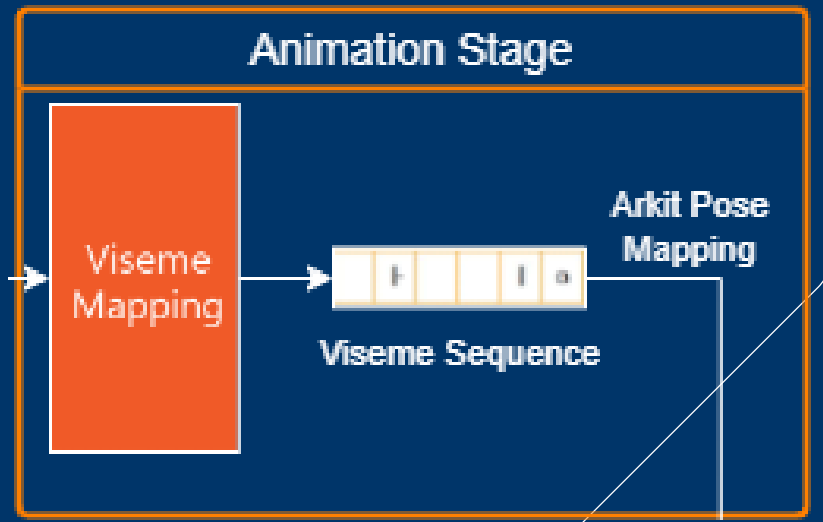
- 오디오로부터 Phoneme Sequence 추출



Animation Stage

1. Phoneme 길이에 따른 처리

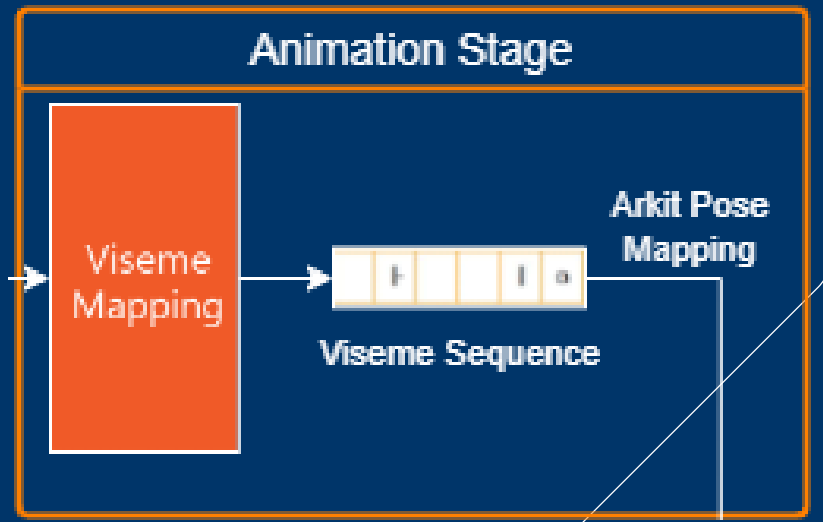
- 길이가 한 프레임인 Phoneme은 이전 프레임에 해당하는 Viseme으로 대체



Animation Stage

2. Phoneme 유형에 따른 Mapping

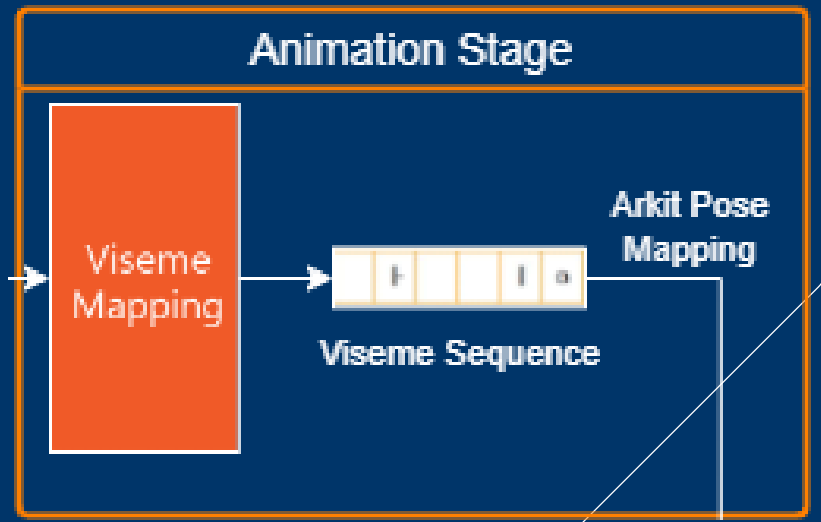
- SIL: 무음 구간은 입 모양을 닫는 Neutral Viseme에 매핑합니다.
- 모음 : 모음은 단모음과 이중 모음으로 나누어 처리합니다.
    - 단모음: 단일 Viseme으로 매핑합니다.
    - 이중 모음: 길이를 절반으로 나누어 두 개의 Viseme으로 매핑합니다.



Animation Stage

2. Phoneme 유형에 따른 Mapping

- 자음 : 자음은 기본적으로 가까운 타이밍의 모음 입 모양에 매핑합니다.
- 예외 1: 치경음('D', 'DH', 'S', 'SH', 'T', 'TH', 'N', 'CH', 'JH', 'Z', 'ZH')
    - 치찰음(s z J C S Z)은 턱을 크게 좁힙니다(예: '체스'의 'C'와 's'는 모두 치아를 서로 가깝게 만듭니다).
- 예외 2: 양순음('m', 'b', 'p', 'f', 'v')은 전후의 viseme이 'Oh', 'Woo'인 경우 PBM2로, 그렇지 않은 경우 PBM1로 매핑합니다.
- 예외 3 : 입술이 많은 viseme(UW, OW, OY, w, S, Z, J, C)은 일찍 시작하고 늦게 끝납니다.
    - 입술이 많은 viseme 은 이웃한 viseme 이 순치 및 양순이 아닐 경우 2프레임 먼저 시작하고 2프레임 늦게 끝나도록 합니다.



Output Stage

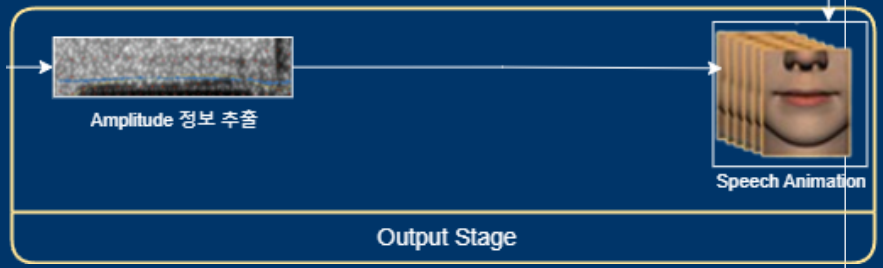
오디오로부터 추출한 Amplitude value를 Viseme pose의 Weight 값에 곱하여 자연스러운 입 모양 생성

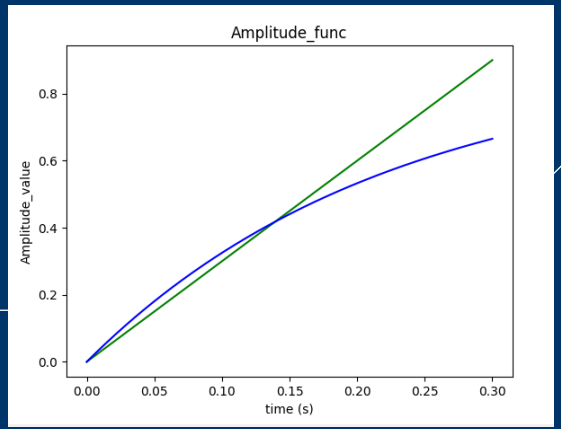
- Amplitude function : (1-5**(-1*np.exp(1)*1.3*amplitude_value))*1.7+0.7
- Exponential function 사용 결정 이유

  1. 주요 고려 범위 (0~0.3) 사이의 두 그래프(Linear, Exponential)를 비교해봤을 때 크게 차이가 없음

     즉, Exponential function을 사용하더라도 Audio에 담긴 정보를 추출할 수 있음

  2. 예상치 못한 Noise가 들어왔을 경우를 대비해 Max값 조정 가능

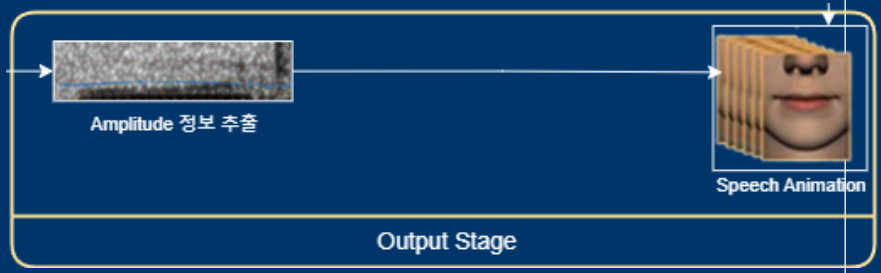


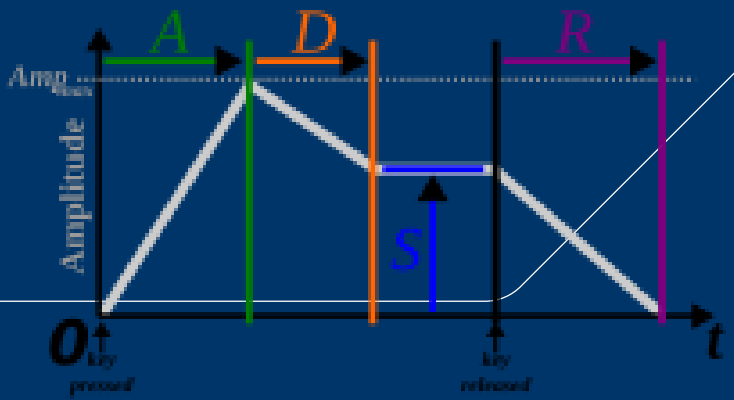


Output Stage

오디오로부터 추출한 Amplitude value를 Viseme pose의 Weight 값에 곱하여 자연스러운 입 모양 생성

- ADSR Curve
  - 자연스러운 립싱크 애니메이션 생성을 위해 Amplitude value가 ADSR Envelope 을 적용
  - 발화 시작 정보와 끝 정보, 한 Viseme의 시작 정보와 끝 정보를 바탕으로 Amplitude value 가 ADSR의 4단계를 따르는 커브 적용





04. Result

Result

- 최대 처리 가능한 오디오 길이 : 5000 프레임
- 처리 속도

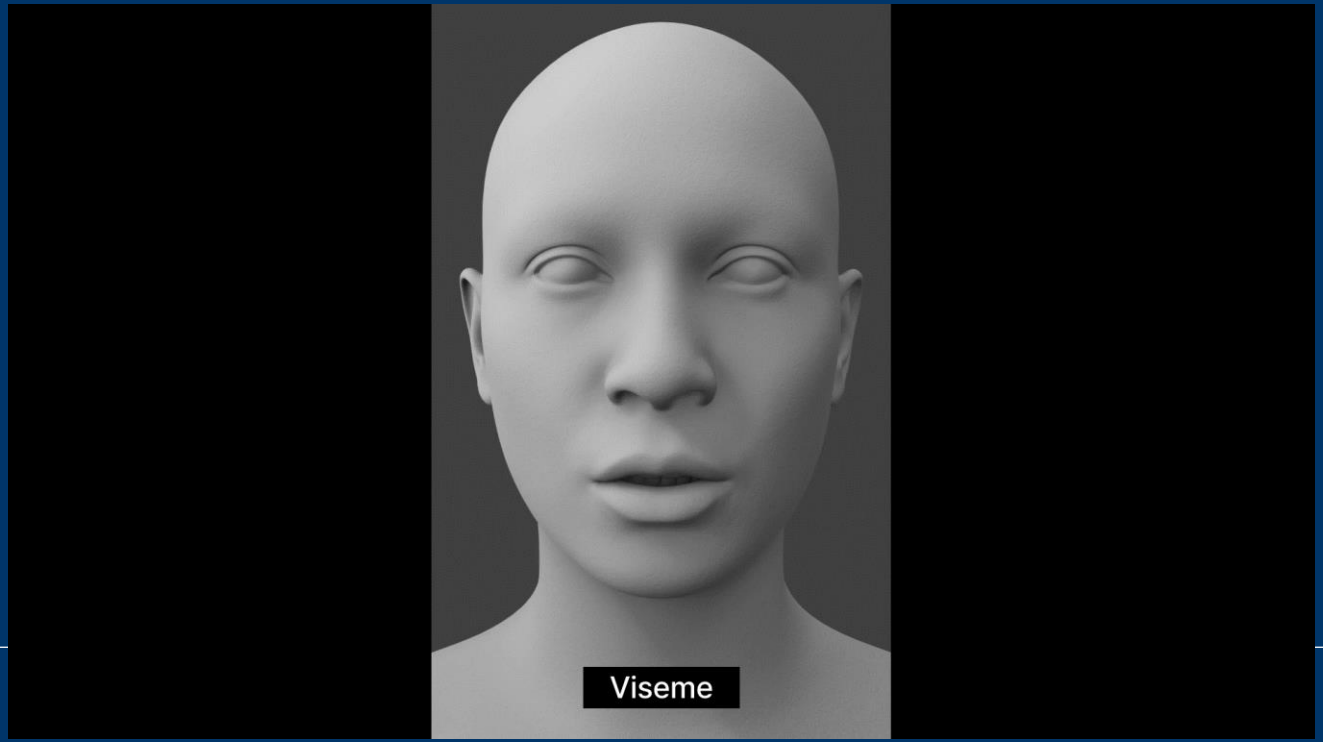
60초 wav(3596 frames) 처리 시간 5.560 sec

Audio load : 1.350 seconds

Viseme model prediction : 4.070 seconds

Rig save : 0.139 seconds

Result Video



Result

- BS weight를 Output으로 사용하고 있기 때문에 추가적인 Solving이 필요 없어 속도가 빠름
- PPG로부터 Phoneme 정보를 추출하여 규칙 기반 Viseme 매핑을 통해 어떤 음성에도 견고한 립싱크 애니메이션 생성이 가능

Limitation

- PPG의 성능에 의존하여 PPG가 정확하지 않을 경우 애니메이션의 정확도 크게 감소
- Viseme에 대응하는 BS weight 확보 실패로 입모양이 다소 부정확함

04
Conclusions

Conclusions

Viseme Lip-sync Model은 PPG로부터 추출한 Phoneme에서 Viseme으로의 규칙 기반 Mapping을 통해 다양한 음성으로 부터 일관된 립싱크 애니메이션을 생성 할 수 있음

REFERENCES

- 한국어 동시조음 모델에 기반한 스피치 애니메이션 생성
- JALI: An Animator-Centric Viseme Model for Expressive Lip Synchronization (toronto.edu)

Part.2 Audio2Blendshapes

01. Introduction

Introduction

* Speech-Driven 3D Facial Animation을 게임에서 얼굴이 부각되는 **메인 캐릭터**에 활용하기 위해서는

성우 연기 음성을 기반으로 **음성 정보와 화자의 특성**이 나타나야 함

* Audio2Blendshape(이하 A2B) 모델은 성우의 음성 데이터를 기반으로 메인 캐릭터에 사용할 수 있는

립싱크 애니메이션 생성이 목표

Goal

- 음성 정보를 바탕으로 ARKit Blendshape 기반 Facial Animation을 생성하는 End-to-end 모델 구성
- 연기 음성에 대한 자연스러운 표현이 가능한 립싱크 애니메이션 생성
- 음성에 담긴 화자의 특성을 반영한 립싱크 애니메이션 생성

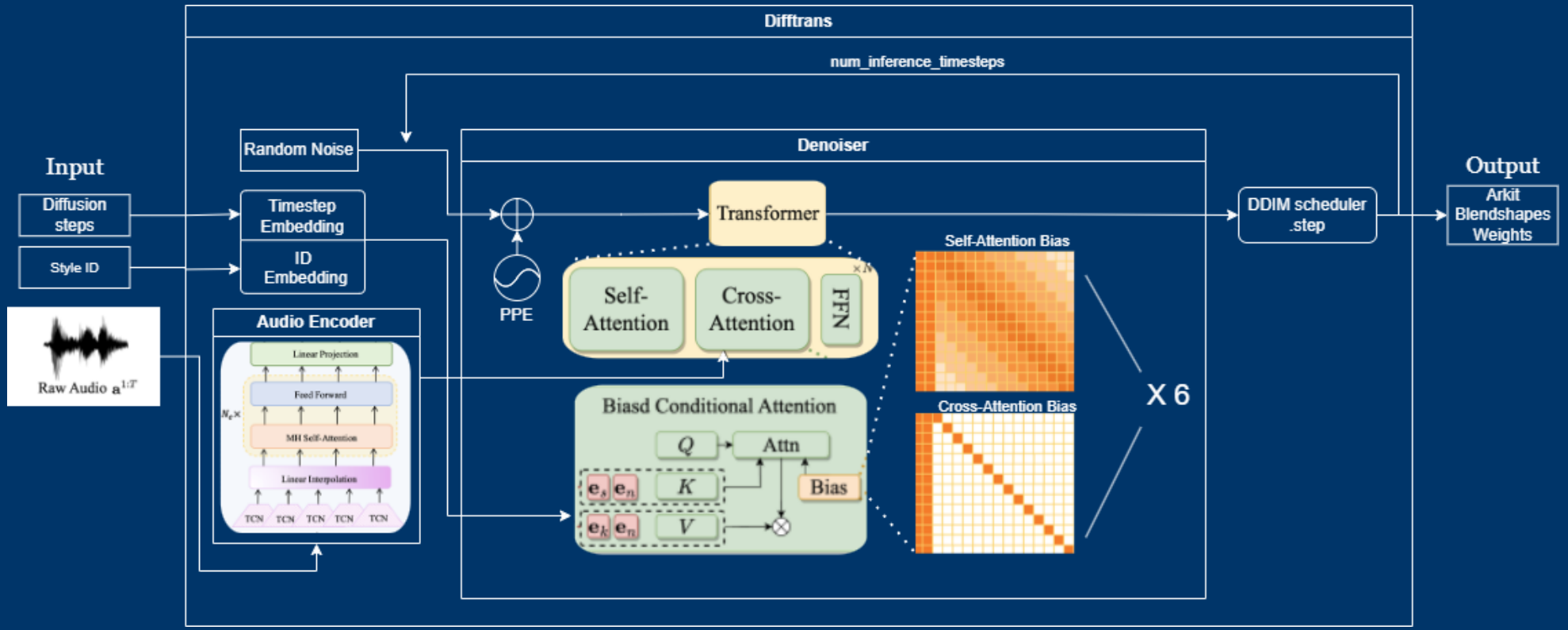
Contribution

- Difftrans는 wav2vec 2.0과 Transformer기반 Diffusion Model를 결합해 성우 음성 특성을 효과적으로 반영한 립싱크 애니메이션 생성 모델
  - Diffusion 모델을 Motion Decoder로 사용하여 Pose 중심 학습을 통해 Fidelity 향상
  - Wav2vec 2.0을 사용하여 다양한 음성에 대한 일반화 능력 향상 & 연기 음성에 대한 표현 가능성 확인
  - Attention with Linear Biases(ALiBi) Method + Periodic Positional Encoding (PPE)을 추가로 도입하여 긴 오디오 시퀀스에 대한 일반화 능력 향상

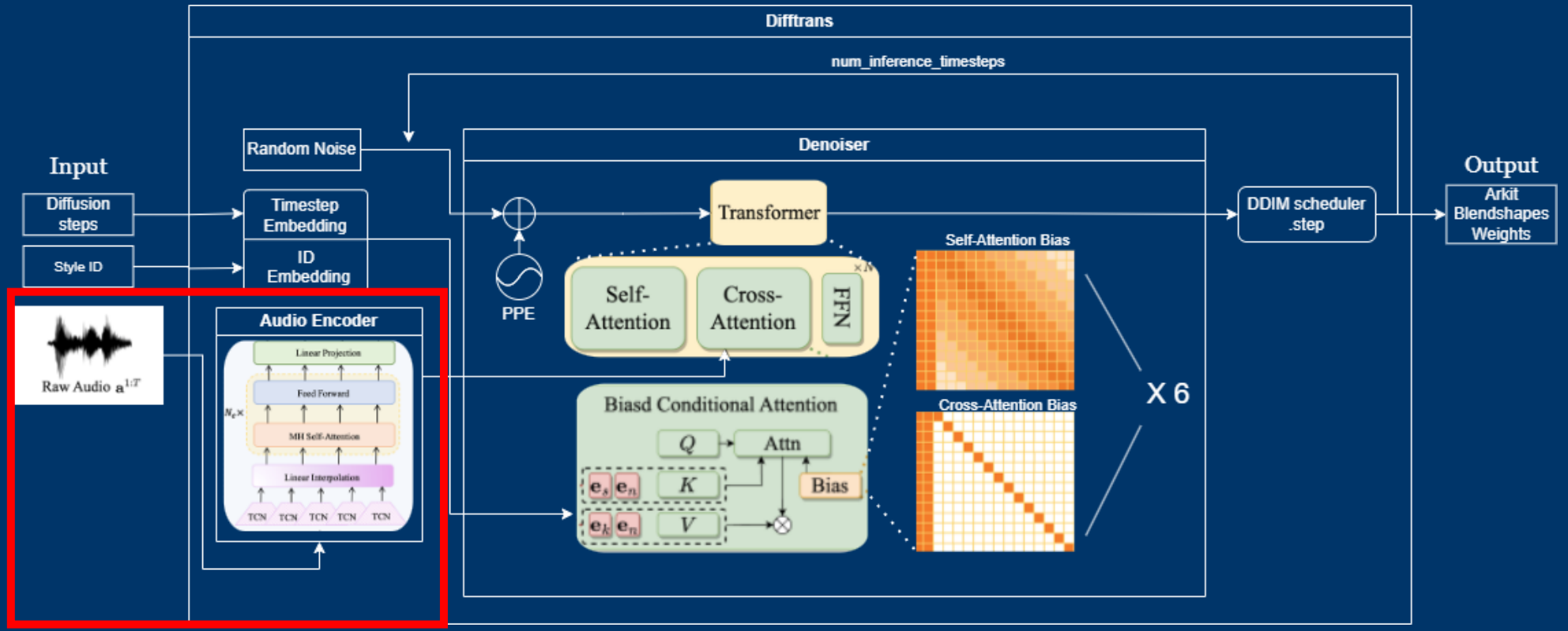
→ 음성에 담긴 **화자의 특성을 반영**한 립싱크 애니메이션을 생성

02. Model

Model structure



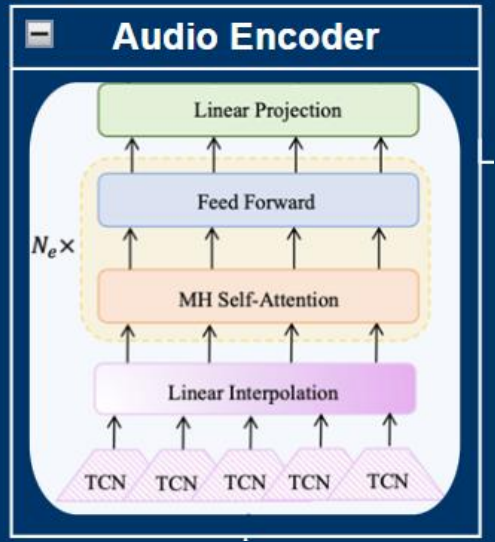
Wav2vec 2.0 audio encoder



Wav2vec 2.0 audio encoder

Wav2vec 2.0 선택 이유

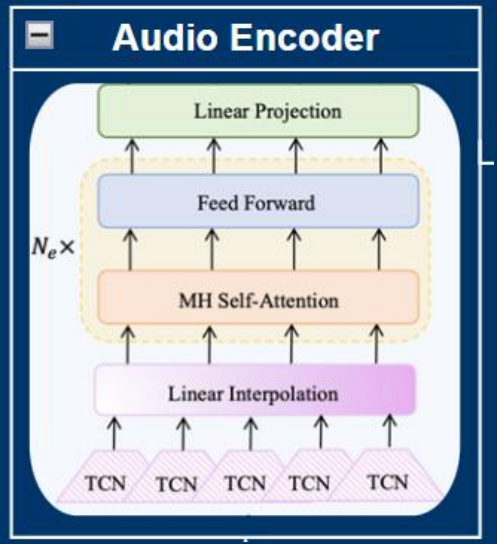
- Audio-Driven 3d Facial Animation 분야의 데이터 부족 문제를 해결
  - Wav2vec 2.0은 사전학습 된 모델을 Finetuning하는 방식으로 적은 양 데이터를 가지고 높은 성능 기대
- 다양한 음성 처리가 가능
  - 다양한 언어, 화자 음성으로 Pretrained 된 모델을 사용, 학습에 사용되지 않은 음성에 대해서도 좋은 성능을 기대
- Script 정보 이외에도 음성에 담긴 화자의 특성을 효과적으로 인코딩 가능



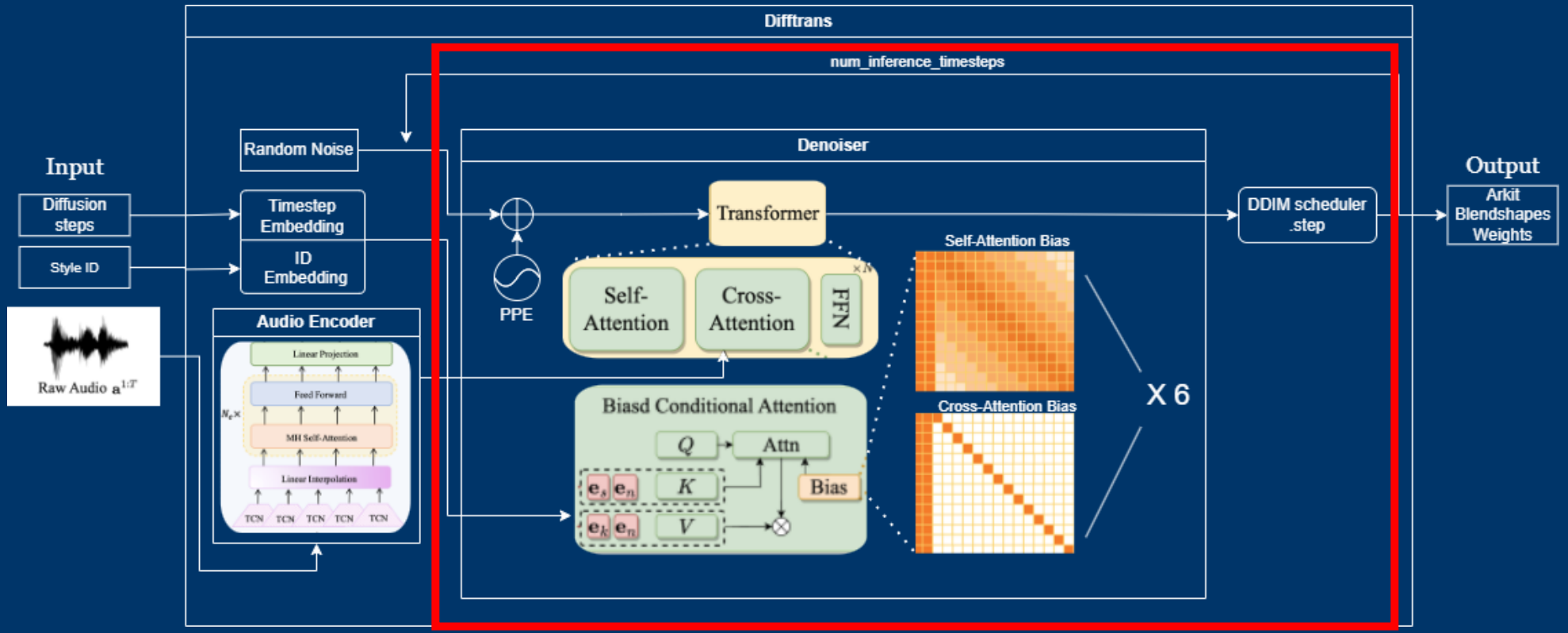
Wav2vec 2.0 audio encoder

Wav2vec 2.0 구조

- Audio Feature Extractor : 여러 개의 Temporal convolutions layer로 구성
  - Raw waveform을 Feature vector로 변환시키는 작업을 수행
- Transformer Encoder : Multi-head self attention과 Feed forward layer로 구성
  - Audio feature vectors를 Contextualized speech representations로 변환시키는 작업을 수행



Transformer기반 Diffuser Decoder (Denoiser)

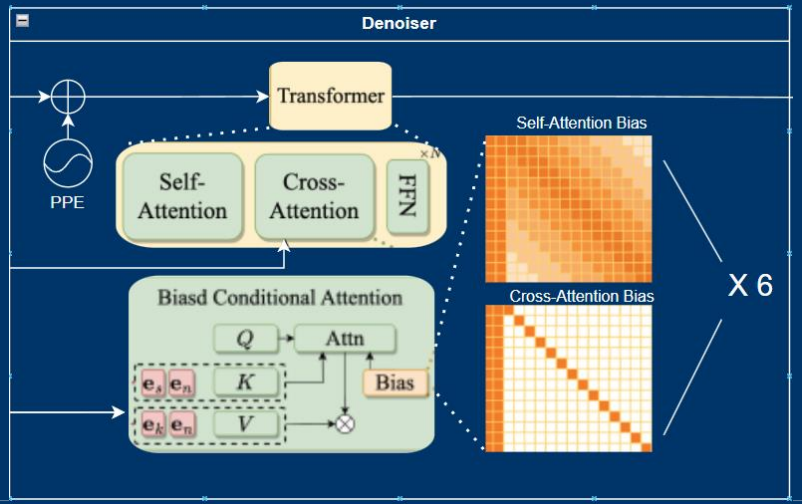


Transformer기반 Diffuser Decoder (Denoiser)

Transformer기반 Diffusion Model 선택 이유

- Diffusion 모델의 '입력에 대해 다양한 결과를 생성 가능하다'는 특성 활용
  - 립싱크 애니메이션 생성에서 웃음 소리나 감정 표현과 같은 비언어적 음성 특성을 반영하는 데 유리

Transformer 기반 Denoiser를 사용하여 음성의 다양한 특성을 반영하면서도 높은 Fidelity를 제공하는 립싱크 애니메이션을 생성

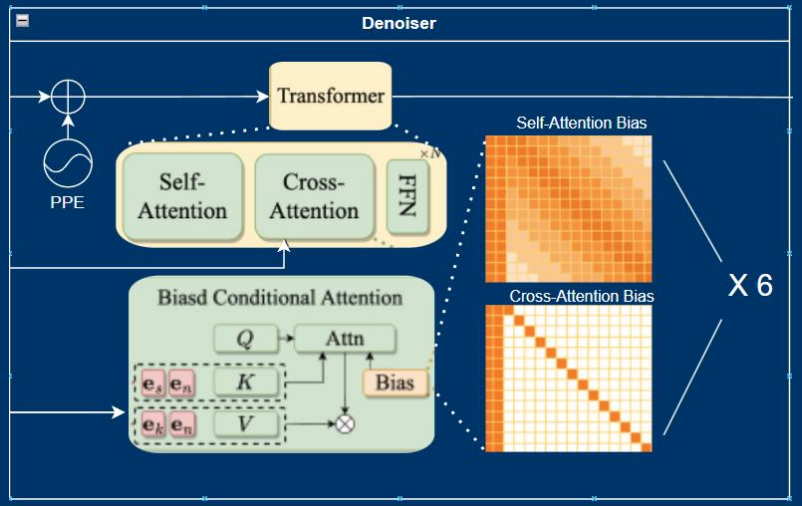


Transformer기반 Diffuser Decoder (Denoiser)

Static Facial expression 문제를 해결하기 위해 긴 오디오 시퀀스에 대한 일반화 능력이 필요

(FaceFormer: Speech-Driven 3D Facial Animation with Transformers)

→ Attention with Linear Biases(ALiBi) method + Periodic positional encoding (PPE)을 도입



Transformer기반 Diffuser Decoder (Denoiser)

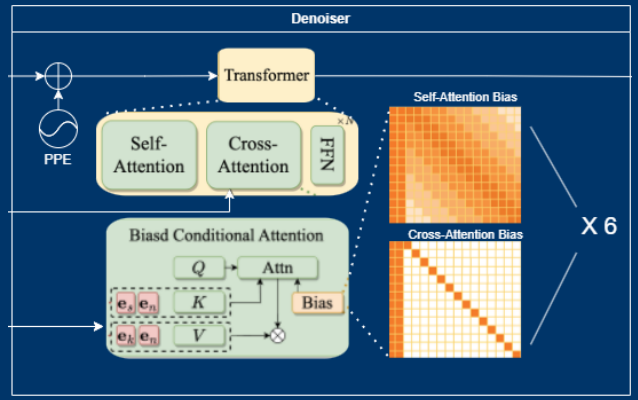
Attention with Linear Biases(ALiBi) method

- Biased causal MH Self Attention : 긴 오디오 시퀀스에 대한 모델의 일반화를 향상시키기 위해 Temporal Bias 사용

$$b_i^s(j) = \begin{cases} 0, & 1 \le j \le 2, \\ \lfloor (i-j)/p \rfloor, & 2 < j \le i, \\ \lfloor (j-i)/p \rfloor, & i < j \le T+2, \end{cases}$$

- Biased Cross-Modal MH Attention : Audio-motion alignment를 위한 Alignment bias를 사용

$$b_i(j) = \begin{cases} 0, & j \in \{1, 2, i\}, \\ -\infty, & otherwise, \end{cases}$$



Transformer기반 Diffuser Decoder (Denoiser)

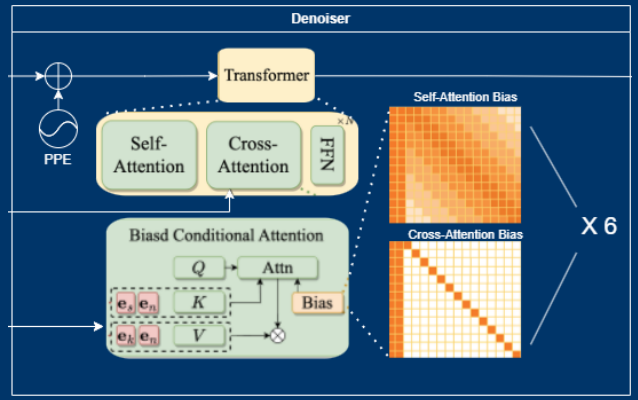
Periodic positional encoding (PPE)

- Periodic positional encoding : 긴 오디오 시퀀스에 대한 모델의 일반화를 향상시키기 위해 도입

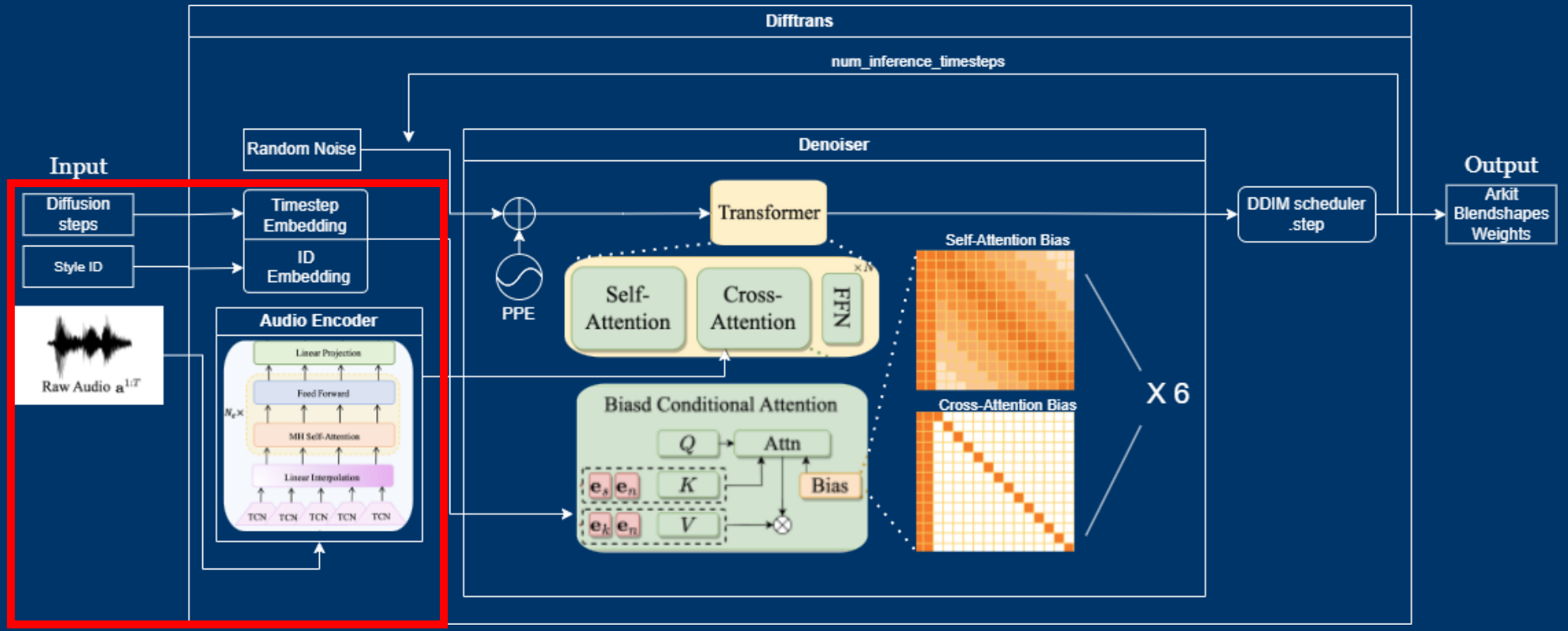
p Period마다 Position information을 주기적으로 주입

$$PPE_{(t,2i)} = \sin(\ (t\ mod\ p/10000^{2i/d})$$

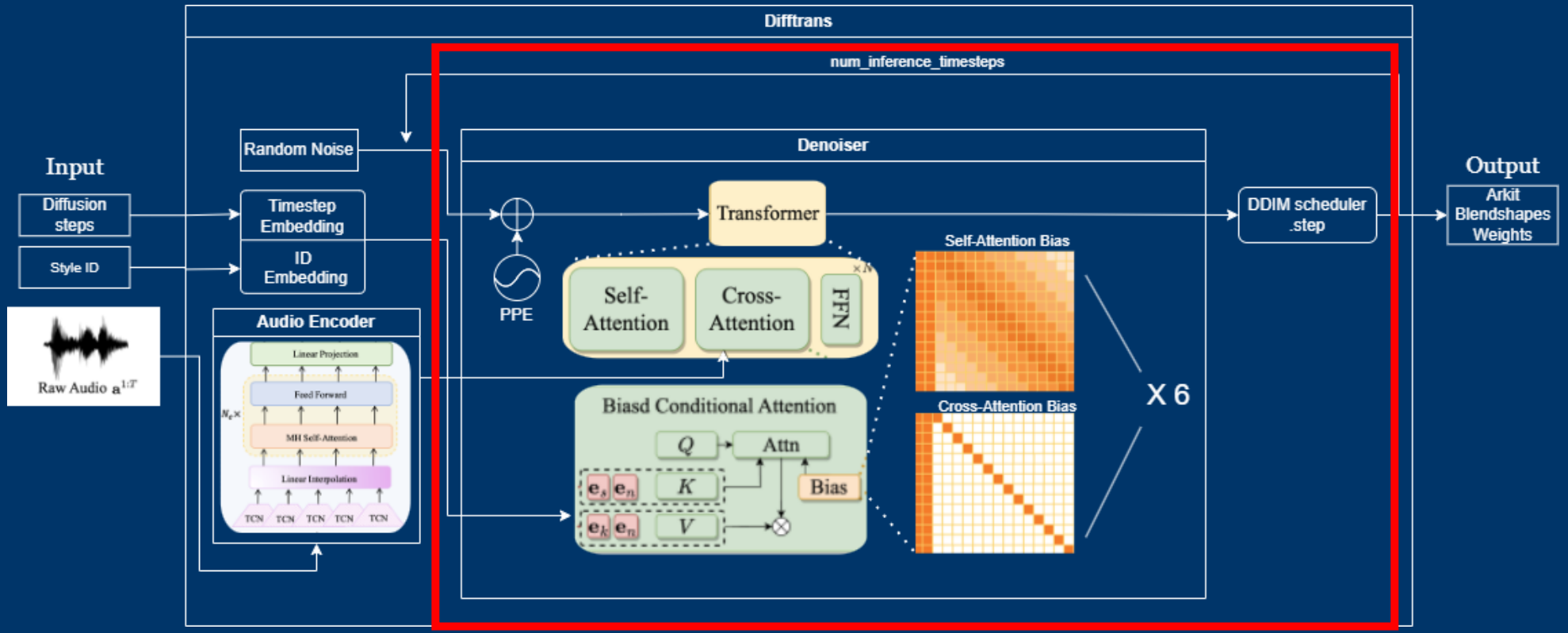
$$PPE_{(t,2i+1)} = \cos(\ (t\ mod\ p/10000^{2i/d})$$



Model structure

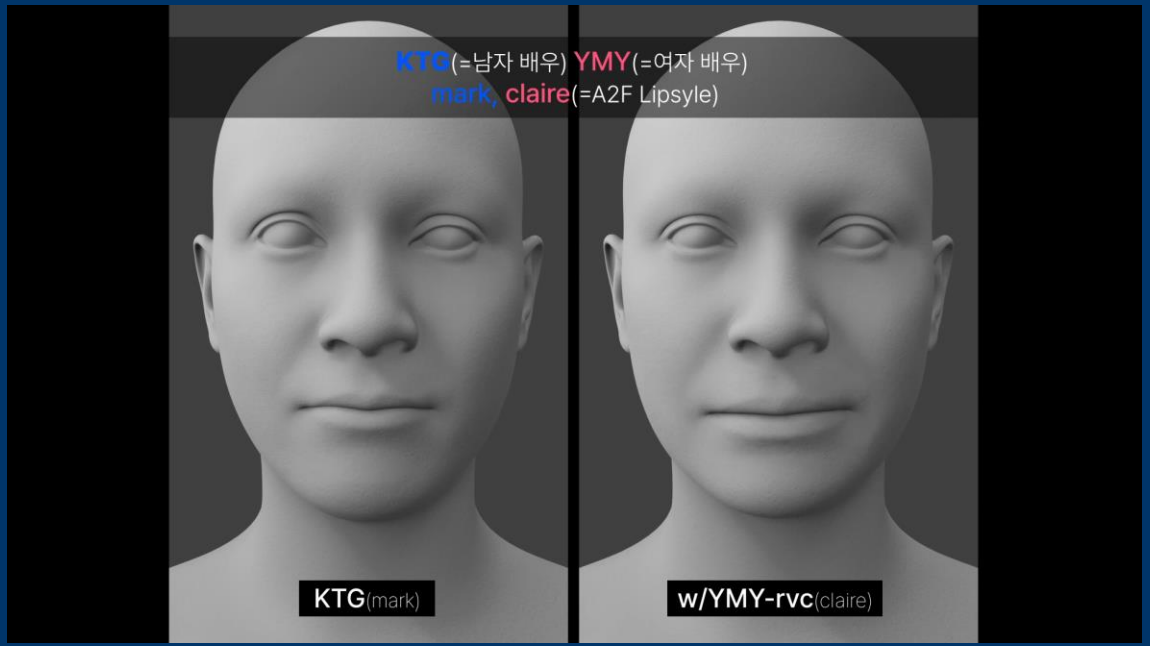


Model structure



학습에 사용된 데이터

- Facial capture 시 녹음된 음성을 사용
- 여성 – YMY(유민영)
  - 257개 스크립트, 약 63분
  - 발화 속도 (보통, 빠르게)
  - 감정 (Neutral, Happy, Surprise, Angry)
- 남성 – KTG(김태규)
  - 734개 스크립트, 약 155분
  - 발화 속도 (보통, 빠르게)
  - 감정 (Neutral, Happy, Sad, Angry)
- 출력 애니메이션 모델
  - Claire / Mark (A2F의 립싱크 스타일)
- Augmentation
  - RVC를 활용해서 YMY <-> KTG 양방향으로 목소리 변환



Data sample

Train

총 400분 가량의 데이터로, 2500 epoch 학습

- Train Data:
  - Female voice : 257 scripts, 63 mins
  - Male voice : 591 scripts, 124 mins
  - Male RVC with Female voice : 745 scripts, 131 mins
  - Female RVC with Male voice : 257 scripts, 63 mins
- Loss:
  - Reconstruction Loss : $L_{rec} = \sum_{n=1}^{N} \sum_{t=1}^{T_n} ||y_{t,n} - \hat{y}_{t,n}||^2$
  - Velocity Loss : $L_{vel} = \sum_{n=1}^{N} \sum_{t=1}^{T_n} ||(y_{t,n} - y_{t-1,n}) - (\hat{y}_{t,n} - \hat{y}_{t-1,n})||^2$
  - Total Loss : $L_{total} = L_{rec} + L_{vel}$

03. Result

Loss

- 기준 : 0.001

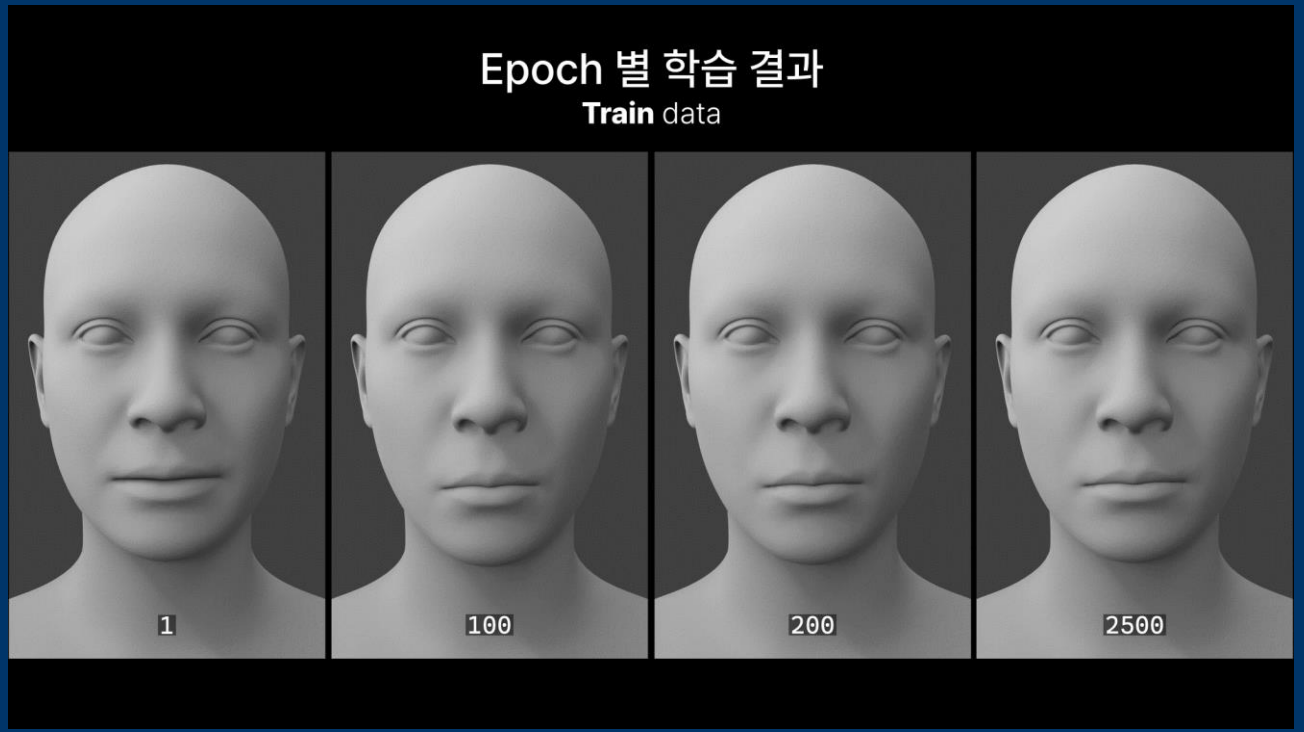

Loss
1.00E-02
9.00E-03
8.00E-03
7.00E-03
6.00E-03
5.00E-03
4.00E-03
3.00E-03
2.00E-03
1.00E-03
0.00E+00
1
501
1001
1501
2001
2501
3001
3501
4001
4501
Total Loss
Velocity Loss
Reconstruction Loss

Inference time

- input : 20 seconds audio
- total load time : **3.91** seconds (vs A2F: **41.45** seconds)
  - model load time : 1.56 seconds
  - wav load time : 1.17 seconds
  - inference load time : 0.93 seconds
  - save load time : 0.24 seconds
- A2F 대비 약 10배 이상 빠른 속도
- BS weight를 output으로 사용하고 있기 때문에 추가적인 Solving이 필요 없어 속도가 빠름

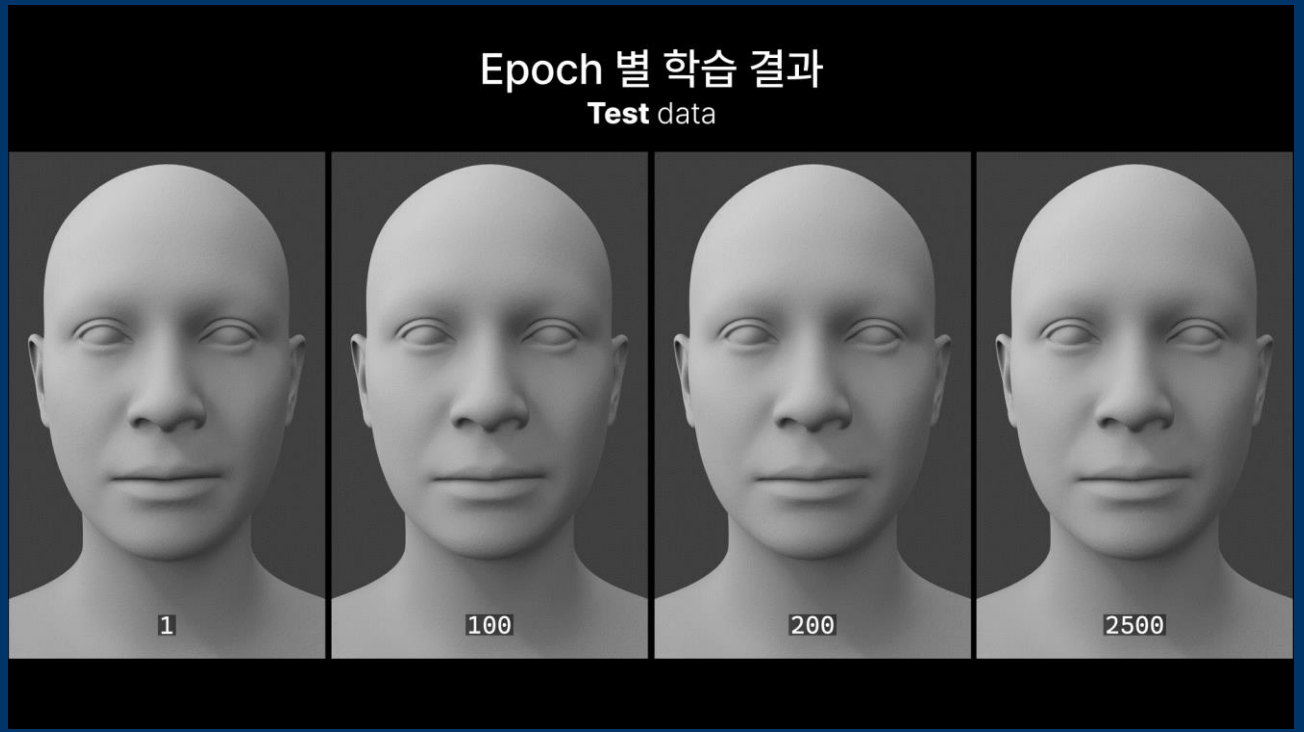
Result Video

- Epoch별 학습 과정



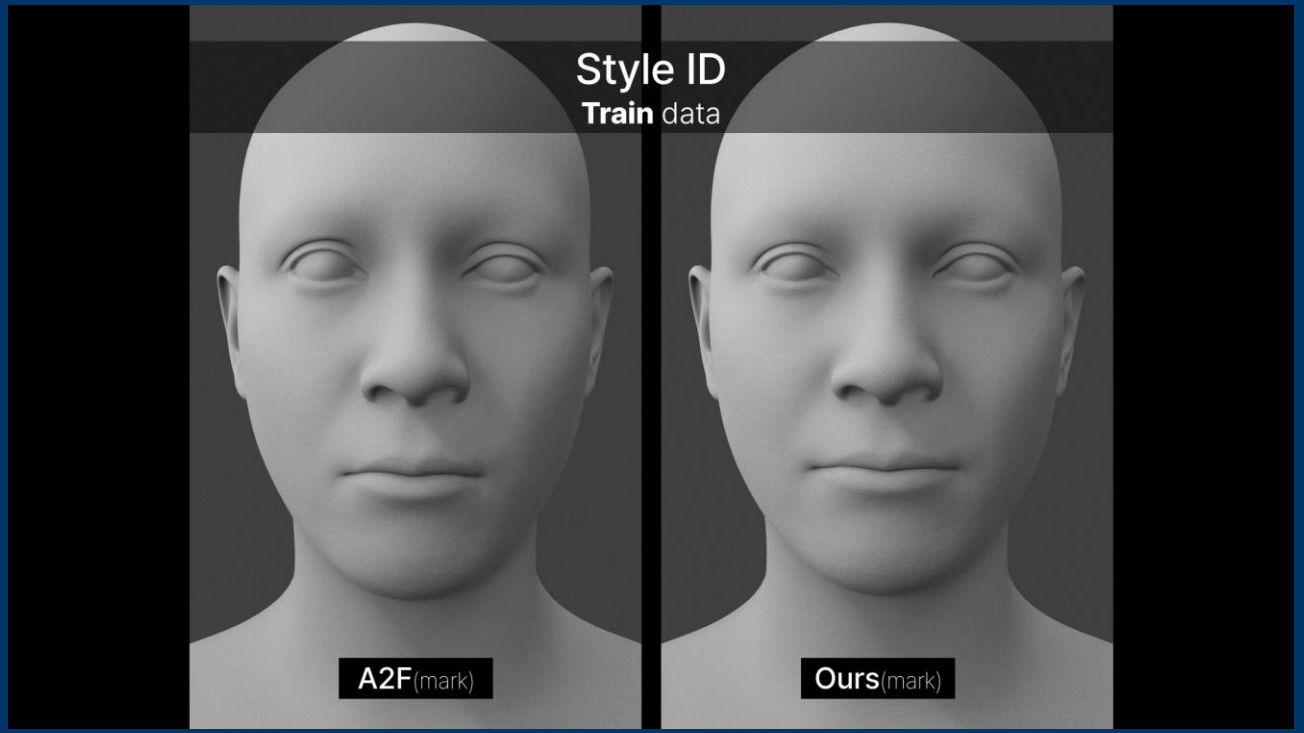
Result Video

- 큰 Pose먼저 학습이 되고 학습이 진행되면서 자연스럽게 이어지는 것을 확인



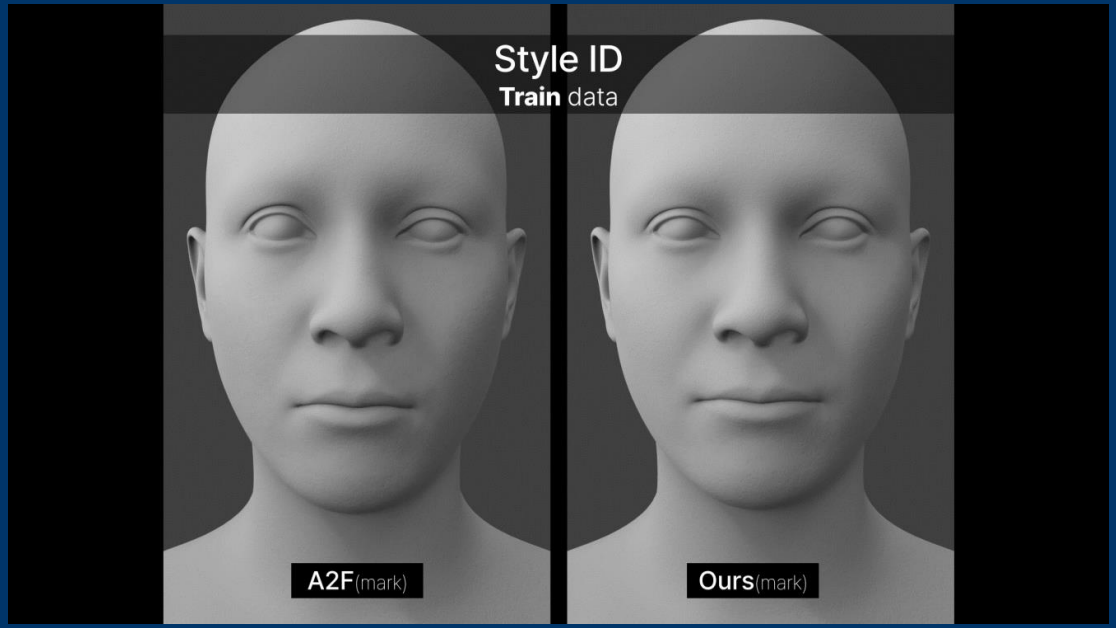
Result Video

- Style ID 성능 검증



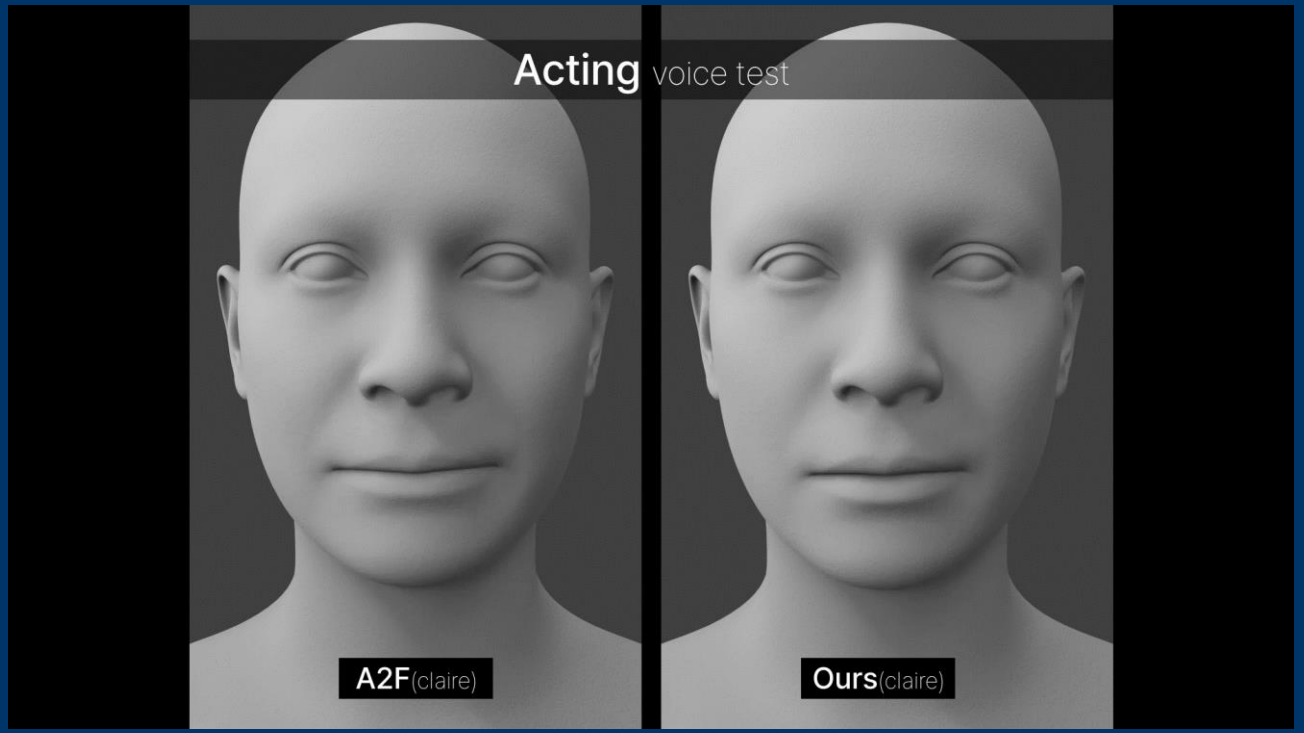
Result Video

- Style ID 성능 검증
  - 2명의 화자 데이터를 사용하여 학습한 결과, 각 화자의 특성이 충분히 반영된 애니메이션을 생성 가능
  - 같은 음성 입력에 대해 Style id를 이용하여 다른 스타일의 립싱크 애니메이션을 생성 가능



Result Video

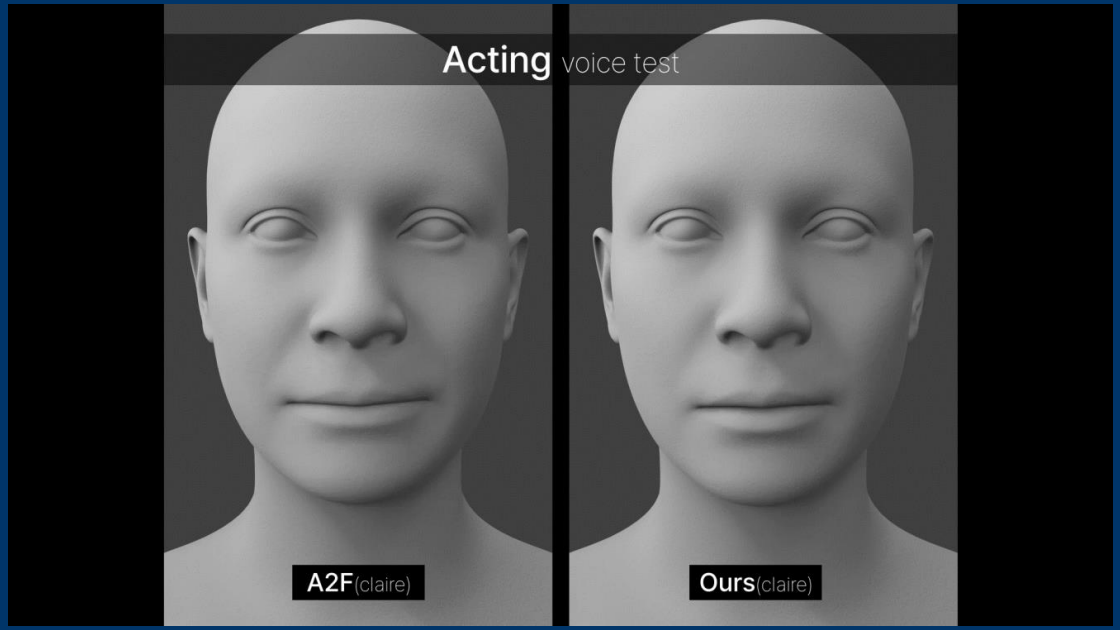
- 연기 음성에 대한 표현 성능 검증



Result Video

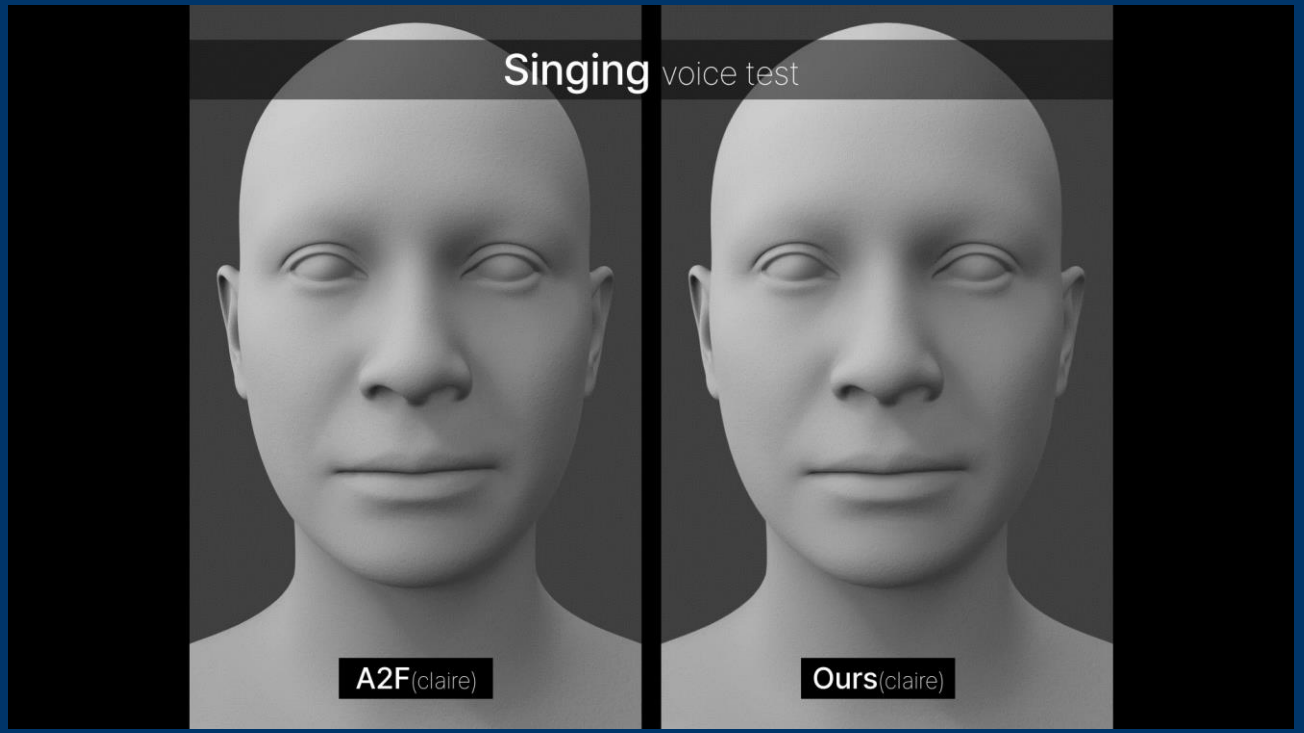
- 연기 음성에 대한 표현 성능 검증
  - 연기 톤 음성에 대해서 자연스러운 애니메이션 생성
  - 비명 소리나 몬스터 음성 정보를 노이즈로 인식하지 않고 처리 가능

  → 비언어적 음성 정보에 대한 표현 가능성



Result Video

- 노래나 음악에서 A2F 보다 노이즈에 강인한 성능을 보임



Limitation

- 2명 이상의 화자 데이터를 사용하여 학습할 경우,
  각 화자의 특징을 유지하면서도 자연스러운 립싱크 애니메이션을 생성할 수 있는지 추가적인 테스트가 필요
- Wav2vec 2.0을 사용하여 웃음소리, 비명소리 와 같은 비언어적 음성에 대한 표현 가능성을 확인했지만
  자연스러운 표현 능력을 향상시키기 위해서 데이터 확보가 필요

04. Conclusion

Conclusions

1. Difftrans는 Wav2vec 2.0과 Diffusion 기반 Transformer를 결합하여 립싱크 애니메이션을 생성하는 End-to-end 모델
2. 음성에 담긴 화자 특성과 성우의 연기에 대한 표현 가능성 확인
3. Style ID를 사용하여 같은 음성에 대해 다양한 스타일의 애니메이션을 생성 가능성 확인

REFERENCES

- DiffSpeaker: Speech-Driven 3D Facial Animation with Diffusion Transformer
- FaceFormer: Speech-Driven 3D Facial Animation with Transformers

End of Document