SAID: SPEECH-DRIVEN BLENDSHAPE FACIAL ANIMATION WITH DIFFUSION

Inkyu Park | Jaewoong Cho

BACKGROUND

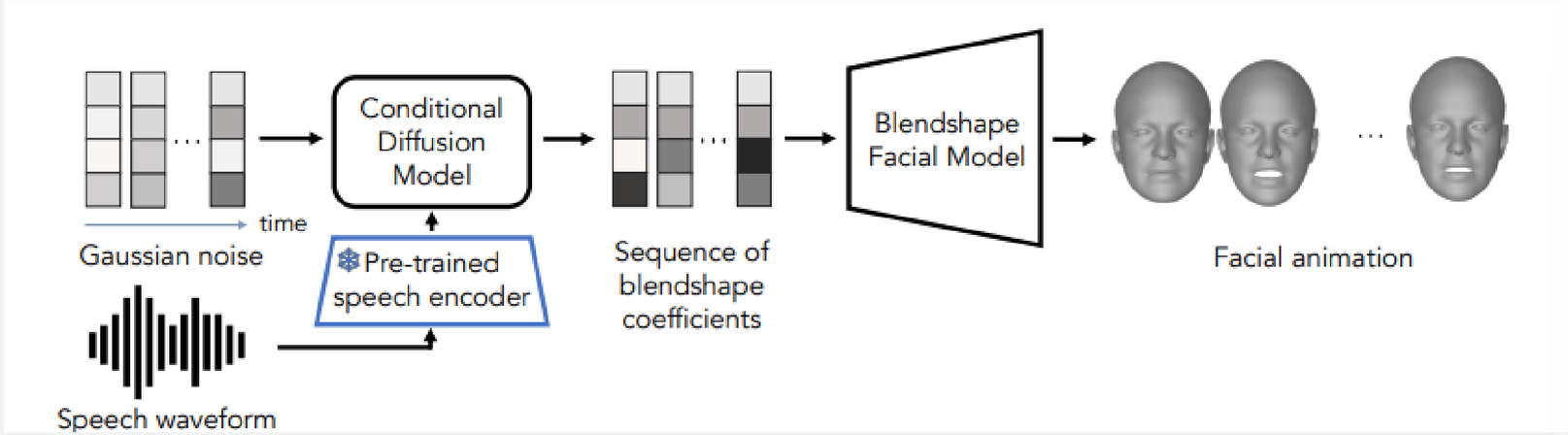
3D 얼굴 애니메이션은 게임, 영화, 가상 현실 등 다양한 분야에서 인간-가상 캐릭터 상호작용을 개선해 왔음

하지만 데이터 얻는 데 시간 및 비용이 많이 소요됨.

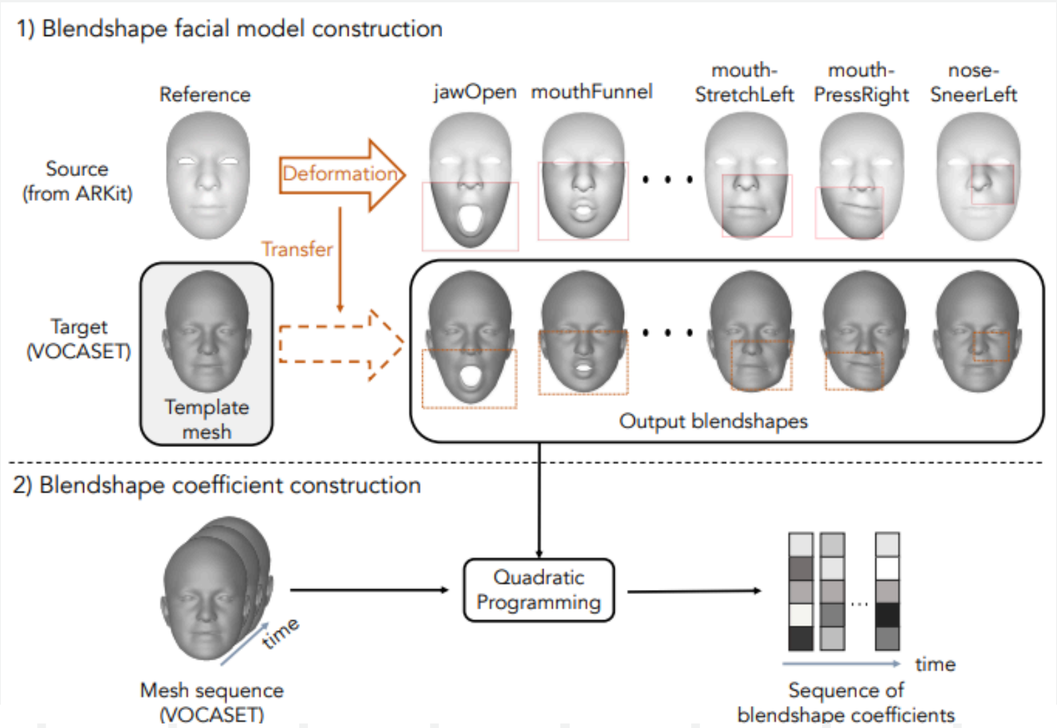
이러한 문제를 해결하기 위해 SAiD(Speech-driven blendshape facial Animation with Diffusion) 모델을 제안함

SAID 모델

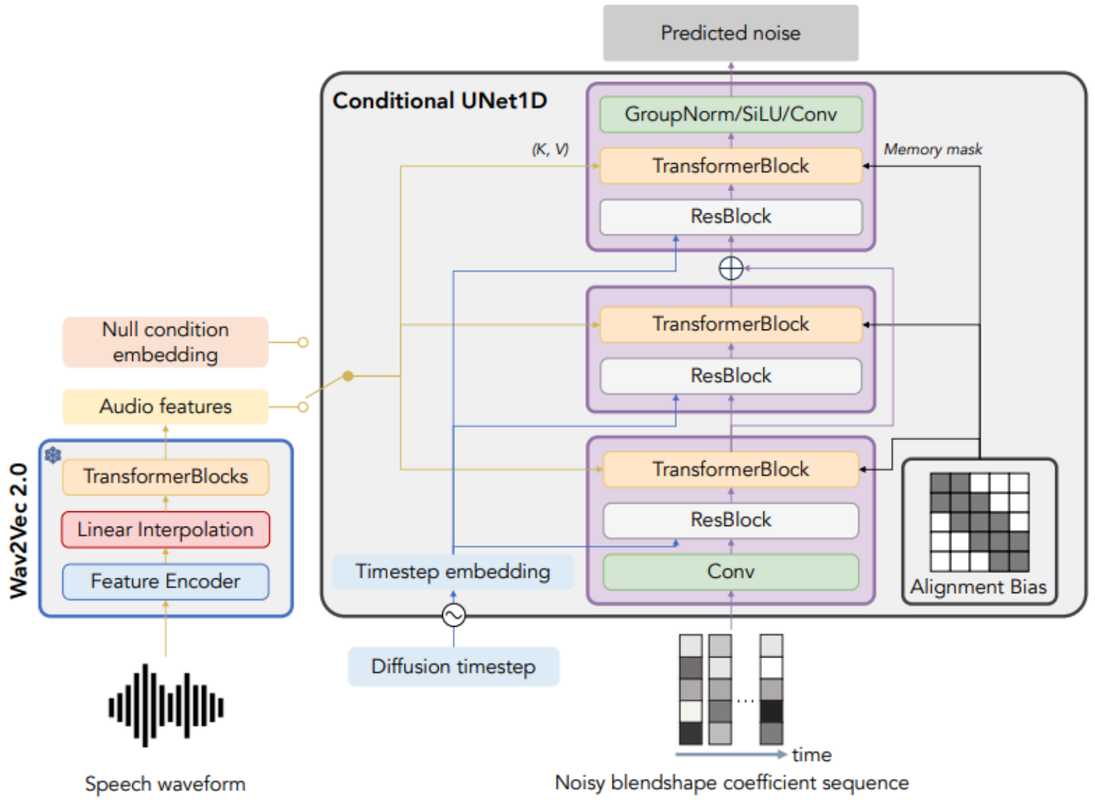
SAiD(Speech-driven blendshape facial Animation with Diffusion)



BLENDVOCA


1) Blendshape facial model construction
Reference
jawOpen
mouthFunnel
mouth-StretchLeft
mouth-PressRight
nose-SneerLeft
Source (from ARKit)
Deformation
Transfer
Target (VOCASET)
Template mesh
Output blendshapes
2) Blendshape coefficient construction
Quadratic Programming
time
Mesh sequence (VOCASET)
time
Sequence of blendshape coefficients

ARCHITECTURE



METHODOLOGY

BLENDVOCA DATASET
12명의 화자로부터, 약 40개의 음성 샘플, blendshape 계수 쌍 구성

MINI BATCH
각 훈련 단계마다 미니배치 크기는 8로 설정

DATA AUGMENTATION
음성 파형을 랜덤으로 1/60초 이동시키거나, 대칭 블렌드쉐입 계수를 서로 교환

TRAINING ENVIRONMENT
NVIDIA A100 (40GB) GPU에서 50,000 에폭 동안 AdamW 옵티마이저를 사용하여 훈련을 진행함.
학습률은 10^-5로 설정되었으며, 가중치 감쇠는 10^-2로 설정함

EXPERIMENTS

기본 모델 성능 평가

SAiD 모델의 성능을 기존의 다양한 모델들과 비교하여 평가

편집 실험

애니메이션 편집 과정에서 모델의 성능을 평가

아블레이션 연구

모델의 각 구성 요소가 성능에 미치는 영향을 평가

RESULT

BASELINE MODELS

End2End AU speech: 음성 스펙트로그램 이용해 blendshape 계수 예측
VOCA: 얼굴 모션의 메쉬 시퀀스를 직접 생성한 모델
MeshTalk: 음성과 관련/비관련 얼굴 특징을 분리하는 모델
FaceFormer: 트랜스포머 기반 자동회귀 모델
CodeTalker: 코드를 사용해 애니메이션 생성 작업을 수행하는 모델

EVALUATION METRICS

AV Offest/Confidence: 입술 움직임과 음성의 동기화 수준
Multimodality: 다양한 입술의 움직임
Frechet Distance, FD, Wasserstein Inception Distance, WInD
: 실제 데이터와 생성된 데이터 간의 유사성

RESULT

SAiD 모델은 입술 움직임의 다양성, 동기화, 부드러움 등 여러 평가 지표에서 기존 모델보다 우수한 성능을 보였음.

특히, AV offset/confidence, multimodality, 프레셰 거리(FD) 측정에서 가장 좋은 성능을 보였음

Methods	AV Offset →	AV Confidence ↑	Multimodality ↑	FD ↓	WInD ↓
Ground-Truth	−1.038	4.874	N/A	0.000	$1.120^{\pm 0.450}$
SAiD (Ours)	**−1.025**	**5.575**	**3.817**	**6.791**	$\underline{10.344}^{\pm 0.127}$
end2end_AU_speech [37]	0.785	0.887	0.000	12.307	$14.070^{\pm 0.076}$
VOCA [11] + QP	$\underline{-0.891}$	3.117	$\underline{2.899}$	45.555	$52.403^{\pm 0.402}$
MeshTalk [41] + QP	−1.532	4.425	0.000	11.106	$13.161^{\pm 0.046}$
FaceFormer [17] + QP	−0.723	$\underline{5.346}$	2.490	10.265	$14.102^{\pm 0.080}$
CodeTalker [58] + QP	−1.476	5.256	1.407	$\underline{6.862}$	$\mathbf{9.813}^{\pm \mathbf{0.067}}$

MOTION EDITING

MOTION IN-BETWEENING	시작과 끝 블렌드쉐입 계수를 고정하고 중간 값을 생성
MOTION GENERATION	특정 블렌드쉐입 계수를 고정하고 나머지 블렌드쉐입 계수를 생성.

ABLATION STUDY

ABSOLUTE ERROR
절대 오차를 사용하면 FD와 WInD 성능 향상, 인지적 정확성 유지에 효과적

VELOCITY LOSS
속도 손실을 적용하지 않은 경우가 결과를 안정적으로 만듦

ALIGNMENT BIAS
정렬 편향을 사용한 결과 평가 지표의 성능 크게 향상

FREEZING THE SPEECH ENCODER

음성 인코더를 미세 조정한 결과 제한된 데이터셋으로 인한
과적합 이슈 발생

CONCLUSION

SAiD는 음성 신호를 기반으로 한 3D 얼굴 애니메이션 생성에 있어
높은 성능을 보여줌.
또한, 애니메이션 편집 과정에서도 우수한 성능을 발휘하여 다양한 애플리케이션에서 활용될 수 있음.
이 논문은 SAiD 모델과 BlendVOCA 데이터셋을 소개하며, 이를 통해 음성 구동 3D 얼굴 애니메이션의 성능을 향상시키는 방법을 제안

THANK YOU