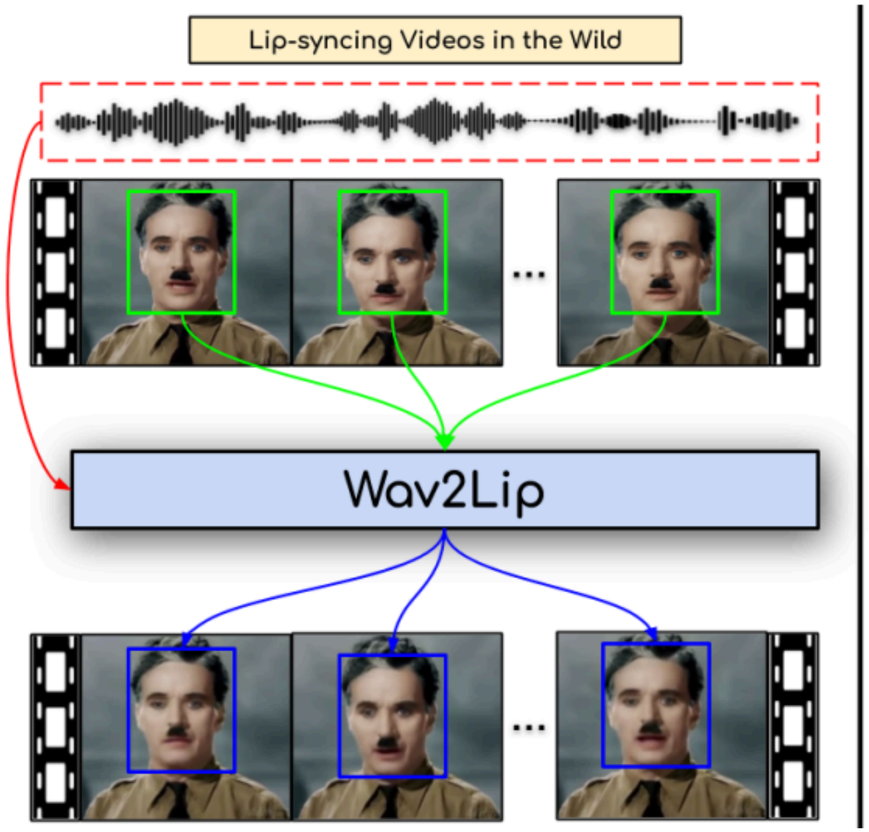
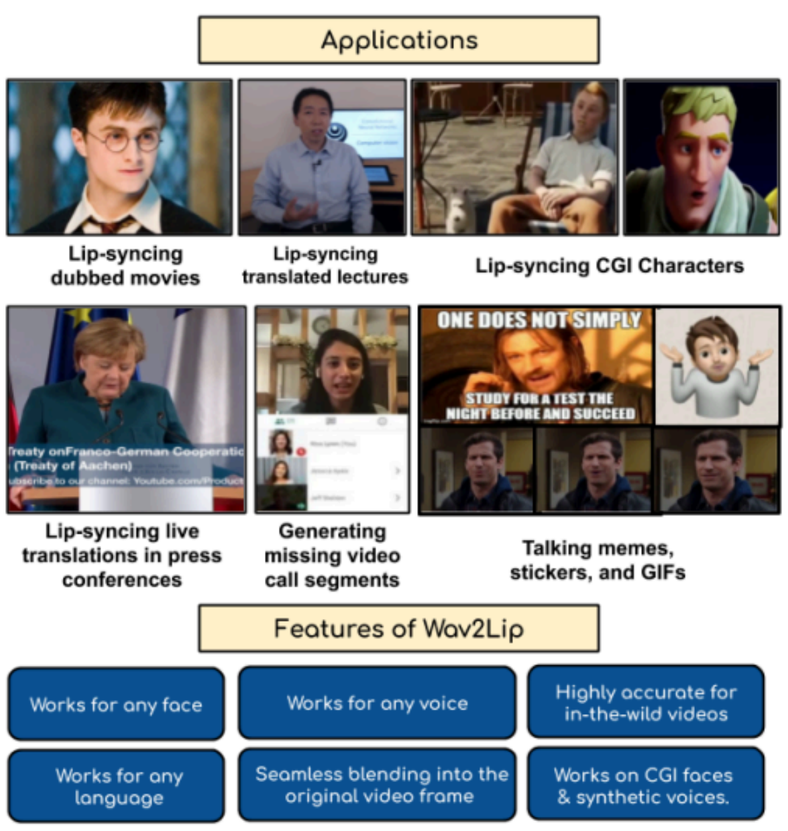
A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild





Presentated by 이혜민

Introduction

문제 제기

오디오-비주얼 콘텐츠 소비 증가
다양한 언어로 비디오 접근성 필요
현재 모델은 특정 화자나 정적 이미지에 한정

연구 목표

동적이고 제약되지 않은 비디오에서 다양한 정체성과 음성의 입술 동기화
화자 독립적인 모델 개발

해결 방안

강력한 lip Sync 판별기(SyncNet) 도입
새로운 평가 벤치마크 및 지표 제안

결과

제안된 모델은 실제 동기화된 비디오와 유사한 정확도를 달성
광범위한 정량적 평가에서 우수한 성능 입증

의의

새로운 평가 프레임워크와 벤치마크 제공
다양한 응용 분야에서 사용 가능성

Related Work

초기 연구

특정 화자의 데이터를 사용하여 음성 표현에서 입술 랜드마크로의 매핑을 학습.
단일 화자에 대해 높은 품질의 결과를 얻었지만, 새로운 정체성이나 목소리에는 적용하기 어려움

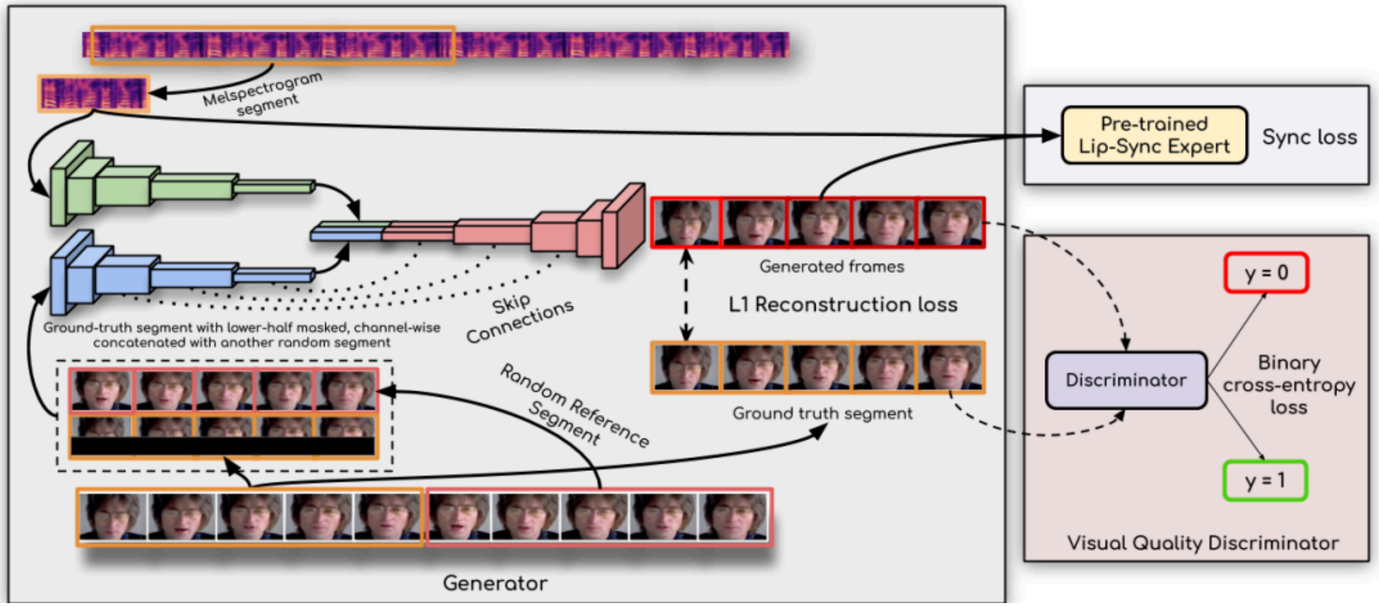
최근 연구

음성 표현에서 직접 이미지를 생성하는 모델들이 등장.
화자 독립적인 모델들이 개발되었으나, 여전히 동적이고 제약되지 않은 비디오에서는 한계가 있음.

주요 한계

기존의 화자 독립적 모델들은 정적 이미지에서는 잘 작동하지만, 다양한 포즈와 조명 변화가 있는 동적 비디오에서는 정확한 입술 동기화가 어려움.

Method



제안된 접근 방식:
사전 훈련된 입술 동기화 판별기(SyncNet)를 사용하여 정확한 입술 동기화를 달성.
입술 동기화 오류를 정확하게 감지하고, 이를 통해 모델이 일관되고 현실적인 입술 움직임을 생성할 수 있도록 함.

기술적 구현:
재구성 손실: 모델이 입력 데이터를 출력으로 재구성할 때의 차이를 최소화.
입술 동기화 판별기: 사전 훈련된 SyncNet을 사용하여 동기화 오류를 판별.
시각적 품질 판별기: 생성된 이미지의 시각적 품질을 평가하여 전체적인 품질을 개선

Experiment

dataset

LRS2, LRW, LRS3 데이터셋을 사용하여 모델을 평가
.

각 데이터셋은 다양한 환경에서 촬영된 비디오 클립을 포함하여, 모델의 일반화 능력을 검증할 수 있도록 구성됨.

Evaluation

LSE-D (Lip-Sync Error - Distance):입술과 음성간의 거리 오차
LSE-C (Lip-Sync Error - Confidence): 입술과 음성 간의 동기화 정도
FID (Fréchet Inception Distance):이미지 품질 평가

Model

Speech2Vid: 이전의 화자 독립적 모델로, 입술 동기화 정확도와 시각적 품질을 비교

LipGAN: GAN 설정에서 훈련된 모델로, 입술 동기화와 시각적 품질을 동시에 평가

Procedure

Wav2Lip과 비교 모델들을 동일한 데이터셋에서 학습하고 테스트.
각 모델의 성능을 평가 지표를 사용하여 비교함
주관적 평가를 위해 평가자들을 통해 각 모델의 출력 비디오 평가.

Result

정량적 평가

Wav2Lip 모델은 가장 낮은 LSE-D 값을 기록, 이는 가장 적은 입술 동기화 오류를 의미.
LipGAN과 Speech2Vid에 비해 상당히 개선된 성능을 보임.
LRS2, LRW, LRS3 데이터셋 모두에서 일관된 성능을 나타냄.

Wav2Lip 모델은 가장 높은 LSE-C 값을 기록, 이는 가장 높은 입술 동기화 신뢰도를 의미.
다양한 데이터셋에서 높은 일관성을 보여, 일반화 능력이 우수함을 입증

Wav2Lip + GAN 설정은 가장 낮은 FID 값을 기록, 이는 시각적 품질이 가장 높음을 의미.
단순한 Wav2Lip 모델에 비해 시각적 품질이 개선됨.

정성적 평가

Wav2Lip 모델이 다른 모델들에 비해 더 높은 선호도를 기록.
특히, 시각적 품질과 입술 동기화 정확도에서 높은 평가를 받음.

Conclusion

- 다양한 비디오에서 정확한 입술 동기화를 위한 Wav2Lip 모델 제안.
- 사전 훈련된 SyncNet 판별기를 사용하여 높은 입술 동기화 정확도 달성.
- 입술 동기화 정확도와 시각적 품질을 동시에 개선.
- 새로운 평가 지표 및 벤치마크, 데이터셋 제시