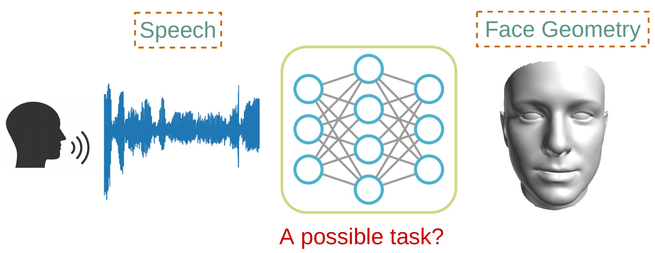
Cross-Modal Perceptionist: Can Face Geometry be Gleaned from Voices?

Cho-Ying Wu, Chin-Cheng Hsu, Ulrich Neumann
University of Southern California

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Motivation

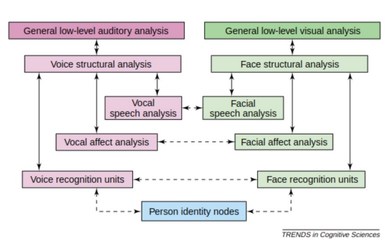
Humans can imagine what a speaker looks like without knowing this person or seeing his face.



Supervised Learning: Learn from paired Voice and 3D face

Unsupervised Learning: Transfer knacks from experts of Image-to-2D and 2D-to-3D

Investigate the correlation between the two modality.



Cognitive science finding
[Trends in Cognitive Sciences, 2004, 2020]

Prior Methods

They use encoder-decoder structure or GAN image synthesis to work on 2D representations.

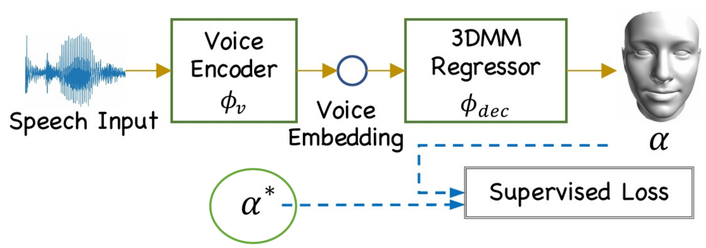
Face image synthesis:

- Irrelevant factors: hair, beards, background, accessories
- Controversial Factor: Race and Ethnicity
- Hard to quantify the reconstruction preciseness

Our study focus: 3D geometry

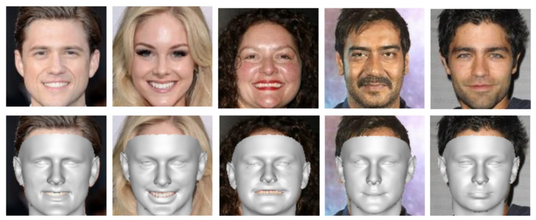
Mesh representation, using BFM Face

Supervised Learning Model



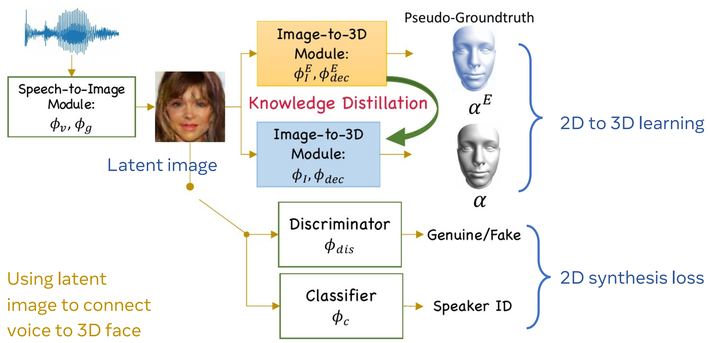
3DMM parameters: controlling face deformation

How to get paired 3D face and voice data?



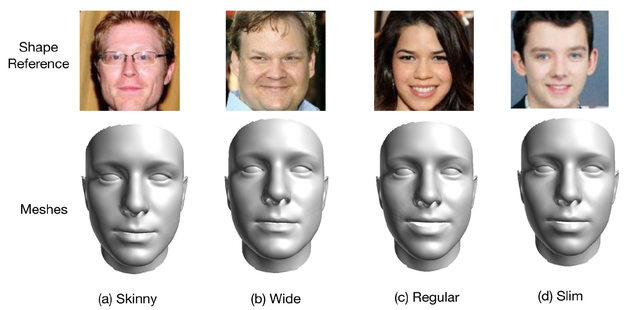
Voxceleb-3D Dataset: Paired voice and 3DMM parameter datasets

Unsupervised Learning Model

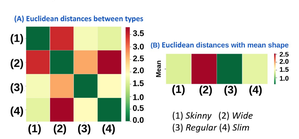


Q1: Is it feasible to predict visually reasonable face meshes from voice?

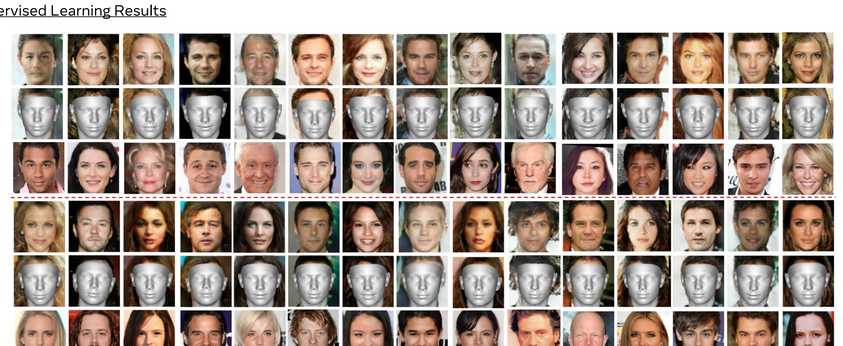
Supervised Learning Results



Affinity Matrix

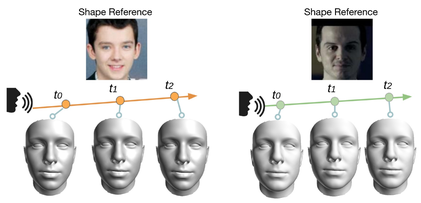


Unsupervised Learning Results

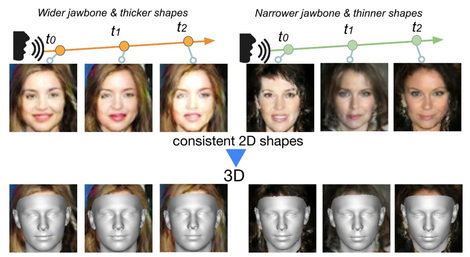


Q2: How stable is the mesh prediction from different utterances of the same person?

Supervised Learning Results

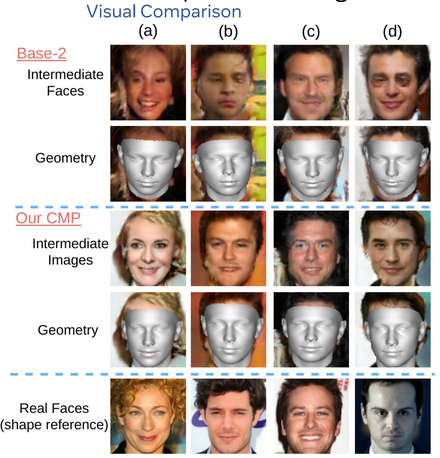


Unsupervised Learning Results



Q3: Comparison with face meshes produced by cascading separately trained speech-to-image and image-to-3D- face methods

Visual Comparison

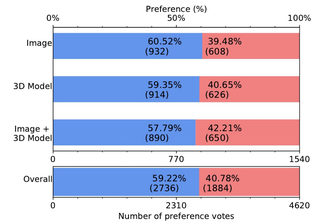


Quantification

ARE	Base-1	Base-2	CMP-supervised	CMP-unsupervised
ER	0.0319	0.0311	**0.0152**	**0.0181**
FR	0.0184	0.0173	0.0186	0.0169
MR	0.0177	0.0173	0.0169	0.0174
CR	0.0562	0.0551	0.0457	0.0480
Mean	0.0311	0.0302	**0.0241**	**0.0251**
Gain	-	0%	**-20.2%**	**-16.9%**

Holistic Registration	Base-1	Base-2	CMP-supervised	CMP- un-supervised
RMSE	1.357	1.348	1.210	1.312

Preference Test



Take a photo to learn more:

SCAN ME