Face processing for visual- and audio-visual speech

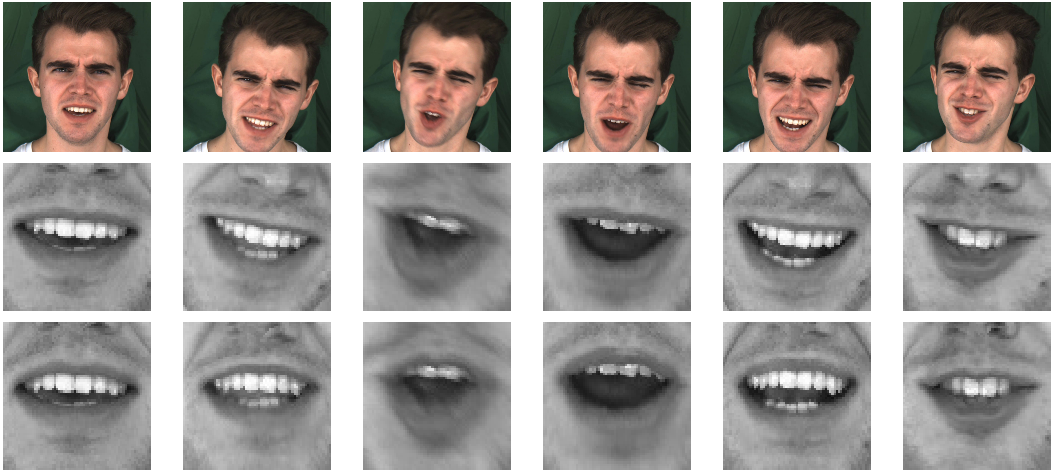
Zhiqi Kang[1], Mostafa Sadeghi[2], Xavier Alameda-Pineda[1], Radu Horaud[1],
Jacob Donley[3] and Anurag Kumar[3]
Inria [1]Grenoble, [2]Nancy, France, and [3]Meta, Redmond, USA

August 23, 2022

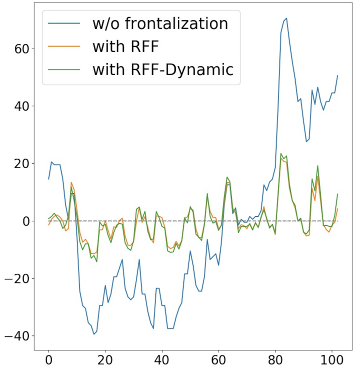
Why *visual* speech?

- Visual perception plays a crucial role in speech communication, e.g. human-to-human and human-to-robot:
  1. Lip, jaw, and tongue movements – non rigid – are controlled by speech production.
  2. Head movements – rigid – play linguistic functions (they mark the structure of the ongoing discourse).
  3. Visual information is not affected by acoustic noise or by competing audio source.
- But:
  1. Non-rigid facial movements cannot be easily separated from rigid head movements, and
  2. Visual information comes with its own caveats, e.g. occluding objects, large variabilities in pronunciation, low resolution, non-verbal lip movements, tongue movements are not observable, etc.

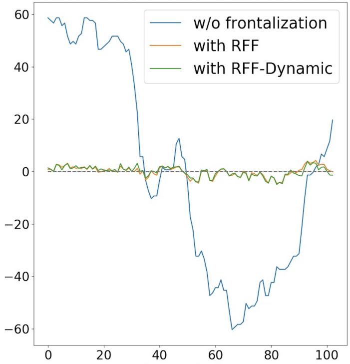
The impact of rigid head motions onto lip movements (i)



The impact of rigid head motions onto lip movements (ii)



Vertical lip motion



Horizontal lip motion

Today's state of affairs

- Until recently, the vast majority of methods combine **noisy speech** with **clean lip motions**, for such tasks as audio-visual speech recognition and speech enhancement;
- Discriminative deep learning techniques have been recently trained with in-the-wild data collections, demonstrating some degree of robustness with respect to visual "noise", e.g. small head movements, low resolution images, self occlusions, etc.
- Nevertheless:
  - **deep lip reading** remains a very difficult task, currently limited to **small vocabulary isolated word recognition**.
  - the vast majority of audio-visual speech processing techniques are discriminatively trained – very large collections of videos are necessary **with associated ground truth**.

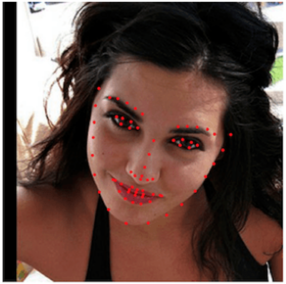
There is a gap between visual- and audio speech recognition

- State of the art lip reading achieves isolated word recognition (IWR) with a small vocabulary: 500-1000 words.
- There are approximatively 170,000 English words in current use, out of 1,000,000.
- Large vocabulary continuous speech recognition (LVCSR) – which is the state of the art in commercially available ASR systems – is out of reach with lip reading.
- Instead we address audio-visual processing, and in particular audio-visual speech enhancement (AVSE)

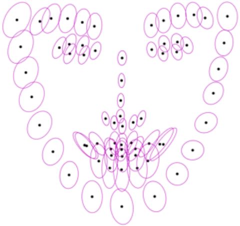
Challenge: How to separate rigid head movements and non-rigid facial movements?

- Face deformation model – 3DMM (3D morphable model),
- Rigid motion model – scale, 3D rotation and translation – 1+3+3 parameters,
- Robust statistical inference of the model parameters,
- Dynamic face frontalization.

Expression-preserving face frontalization

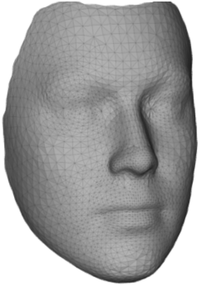


$X_1 \ldots X_N$: 3D facial landmarks



Frontal landmark model:

- Neutral face (means): $Y_1 \ldots Y_N$
- Non-rigid variabilities (covariances)



Deformable face model:

$$V_n = \mathbf{U}_n s + \overline{M}_n$$

Deformable face model

Frontal landmarks are predicted by:

$$\boldsymbol{Y}_n = \mathbf{U}_n \boldsymbol{s} + \overline{\boldsymbol{M}}_n + \boldsymbol{F}_n, \quad \forall n \in \{1 \dots N\}$$

with:

- $\mathbf{U}_n$: reconstruction matrix (learned),
- $\overline{\boldsymbol{M}}_n$: neutral face (learned),
- $\boldsymbol{s}$: low-dimensional face embedding (shape parameters),
- $\boldsymbol{F}_n$: reconstruction error.

Rigid motion model

Frontal landmarks are predicted by:

$$\boldsymbol{Y}_n = \rho \mathbf{R} \boldsymbol{X}_n + \boldsymbol{T} + \boldsymbol{D}_n, \quad \forall n \in \{1 \ldots N\}$$

with:

- ρ: global scale
- $\mathbf{R}$: 3D rotation matrix,
- $\boldsymbol{T}$: 3D translation vector,
- $\boldsymbol{D}_n$: error vector (non-rigid motion, noise, outliers).

Robust estimator: generalized Student-t distribution

$$\boldsymbol{E}_n = \underbrace{\rho \mathbf{R} \boldsymbol{X}_n + \boldsymbol{T}}_{\text{rigid}} - (\underbrace{\mathbf{U}_n \boldsymbol{s} + \overline{\boldsymbol{M}}_n}_{\text{deformable}})$$

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{X}) = -\sum_{n=1}^{N} \log p(\boldsymbol{E}_n; \boldsymbol{\theta}) \tag{1}$$

$$p(\boldsymbol{E}_n; \boldsymbol{\theta}) = \int_0^\infty \mathcal{N}(\boldsymbol{E}_n; 0, \omega_n^{-1} \boldsymbol{\Sigma}) \mathcal{G}(\omega_n; \mu, 1) d\omega_n$$

$$\boldsymbol{\theta} = (\rho, \mathbf{R}, \boldsymbol{T}, \boldsymbol{s}, \boldsymbol{\Sigma}, \mu)$$

Direct minimization of (1) is intractable...

Inference

Initialization

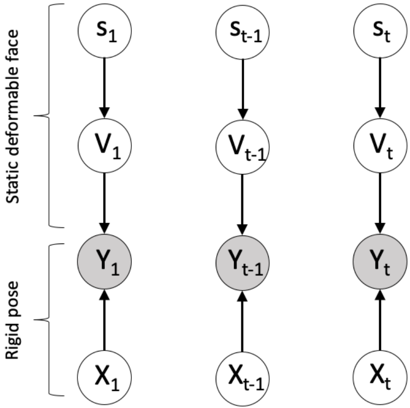
Expectation:

- Evaluate the weight posteriors and the weight means $\overline{w}_{1:N}$

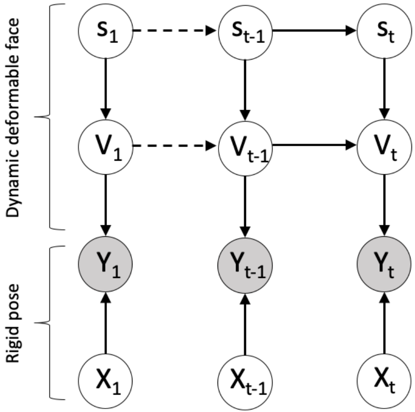
Maximization:

- Estimate the rigid parameters $\rho, \mathbf{R}, \boldsymbol{T}$
- Estimate the non-rigid parameters $\boldsymbol{s}$
- Estimate the pdf parameters $\boldsymbol{\Sigma}, \mu$

Graphical models



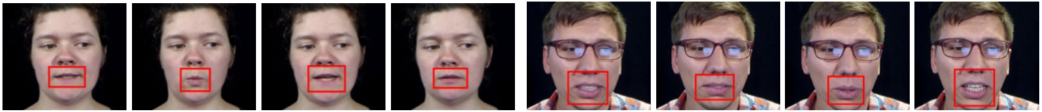
Rigid-and-deformable model



Dynamic deformable model
(doubly latent model $\rightarrow$ Kalman filter equivalence)

Examples from the Oulu dataset

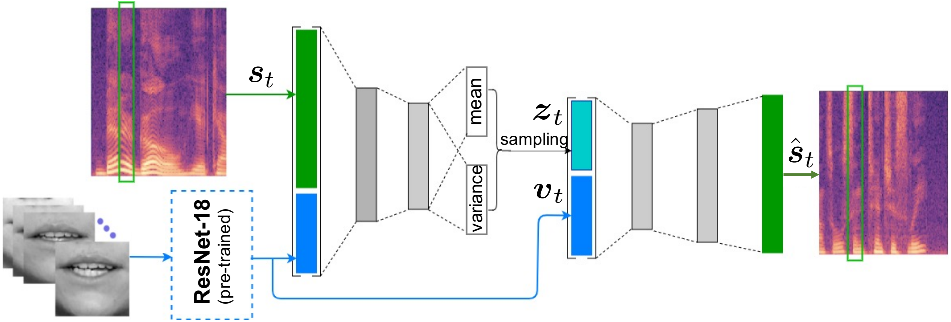
(a) Faces recorded with the 30° camera



(b) Faces recorded with the 0° camera

(c) Proposed (self-occluded regions are displayed in white)

Audio-visual speech enhancement pipeline



For more details please consult [Sadeghi et al 2020, 2021], [Kang 2022].

Speech enhancement results

Measure	STOI $[0,1]\uparrow$					PESQ $[-0.5,4.5]\uparrow$					SI-SDR (dB) $\uparrow$				
SNR (dB)	-10	-5	0	5	10	-10	-5	0	5	10	-10	-5	0	5	10
Noisy audio input	0.40	0.53	0.66	0.78	**0.86**	0.90	1.24	1.67	2.05	2.42	-15.92	-10.62	-5.44	-0.40	4.60
A-VAE Leglaive et al. MLSP'18	0.41	0.56	0.70	**0.79**	0.85	0.93	1.51	2.02	2.43	2.73	-7.01	-0.29	5.08	9.41	12.74
AV-CVAE Sadeghi et al. TASLP'20	0.42	0.57	0.69	**0.79**	0.84	1.02	1.56	2.06	2.42	2.73	-6.96	-0.04	5.01	9.06	12.25
Res-AV-CVAE-DFF	**0.43**	**0.60**	**0.73**	**0.79**	0.85	**1.13**	**1.71**	**2.20**	**2.48**	**2.77**	**-6.35**	**0.28**	**5.87**	**9.42**	**12.77**

Table: Average STOI, PESQ, SI-SDR values.

Examples & paper download:
ICASSP'22:
`https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9746401`
IJCV submission: `https://team.inria.fr/robotlearn/research/facefrontalization-benchmark/`

Conclusions

- We proposed robust face frontalization (RFF) and its dynamic extension (DFF).
- Both RFF and DFF rely on 68 facial landmarks:
  - Sufficient to show that face frontalization improves audio-visual speech performance,
  - Insufficient to really boost the performance of audio-visual speech.
- Future work directions:
  - It is planned to use dense facial features to increase the impact of the dynamic model.
  - *Conversational speech* (CHIME-6 Challenge) may benefit from visual processing: Where is the speaker in the room? Who speaks to whom? Who speaks when? etc.