Lip Reading by Leveraging Hahn Convolutional Neural Networks in Low-resourced Environments

Hicham Hammouchi

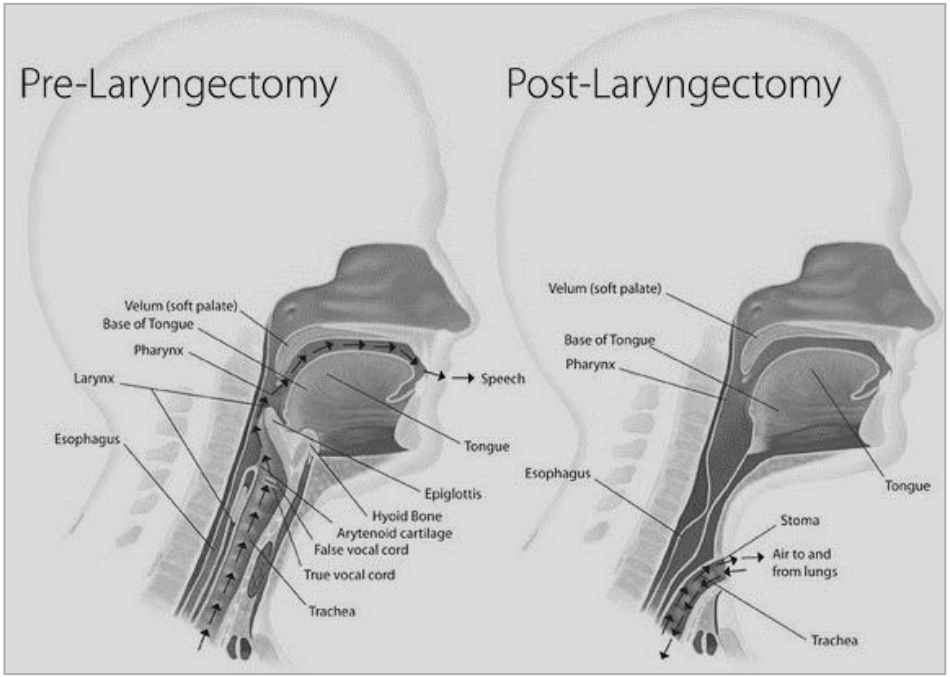
Sidi Mohammed Ben Abdellah University, International University of Rabat

hicham.hammouchi@uir.ac.ma | hi.hammouchi@gmail.com

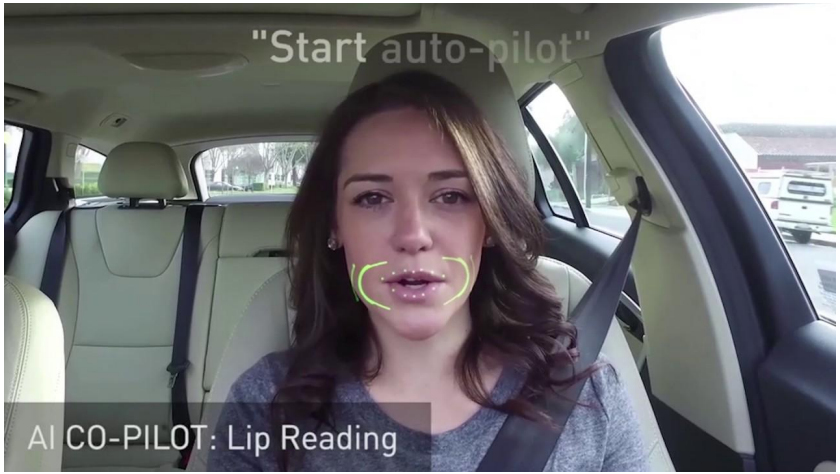
What’s Lip reading?

- Visual Speech Recognition is about understanding what a person is saying by looking at the lips movements
- Lip reading is a hot topic combining two AI fields, Computer Vision and NLP
- Lip reading from dream to reality
  - As AI made many applications and tasks possible especially in computer vision and recognizing things

Why should we build AI Lip Reading Systems?



Source: [NALC Laryngectomy UK]



Source: [Nvidia Drive IX]

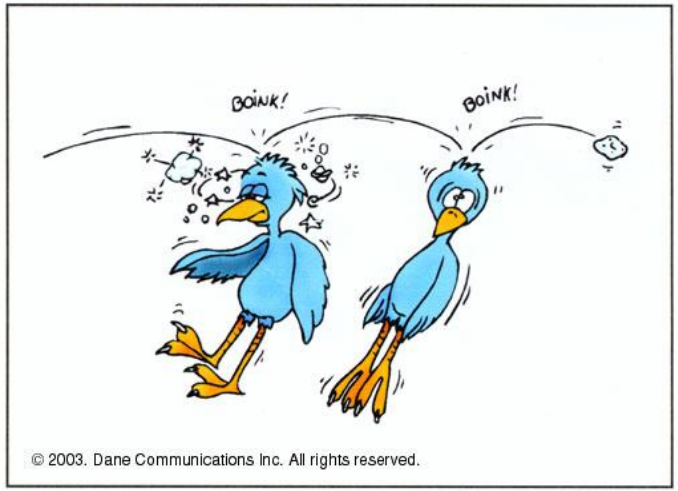
Deep Learning is about depth

Need to go deeper and build deep architectures

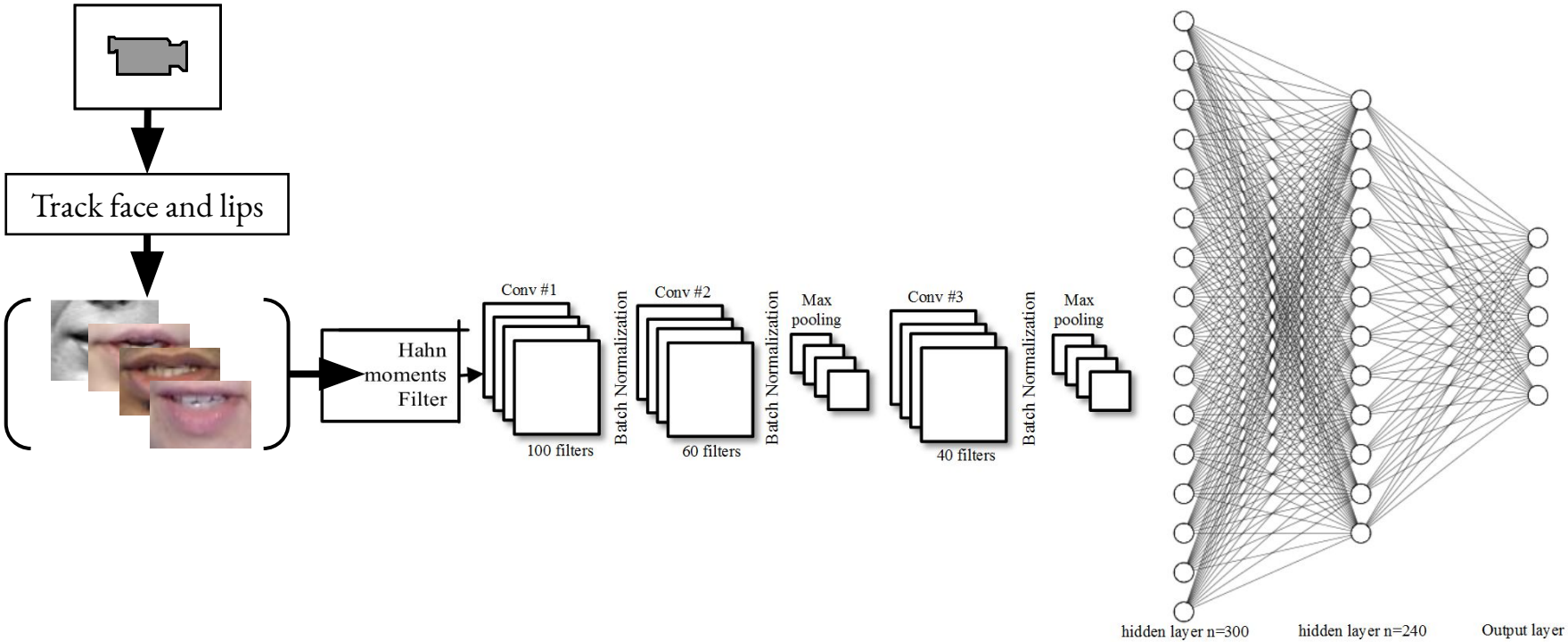
But at the cost of powerful computation resources

What if we stay shallow ?= satisfactory performance guaranteed

- Deal with the high dimensionality in videos with less resources
- Hit two birds with one stone
  - Shallow architecture
  - Satisfactory performance
- Combine Hahn moments and ConvNets in HCNN



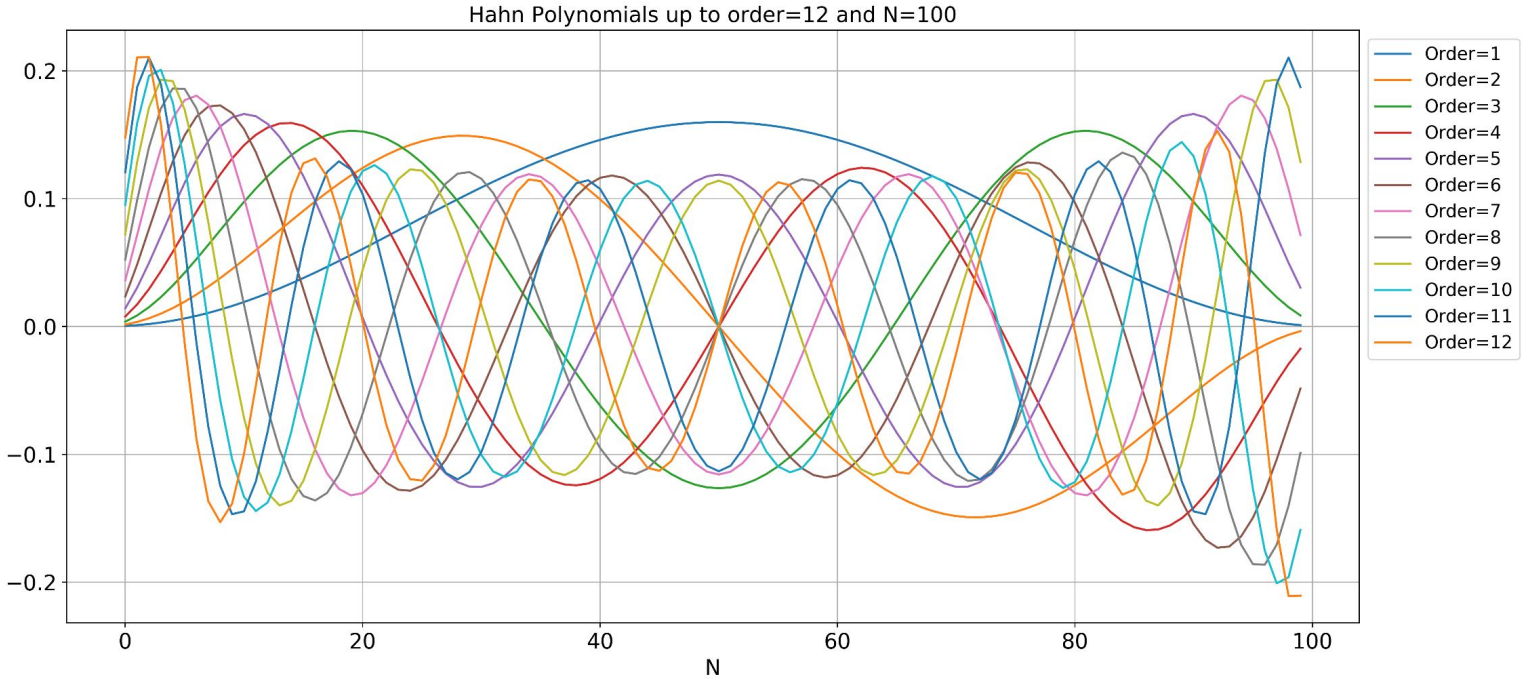
HCNN Architecture



Discrete orthogonal Hahn polynomials

$$h_n^{(\alpha,\beta)}(x,N) = (N+\beta-1)_n(N-1)\sum_{k=0}^{n}(-1)^k\frac{(-n)^k(-x)^k(2N+\alpha+\beta-n-1)^k}{(N+\beta-1)_k(N-1)^k}\frac{1}{k!}, (\alpha,\beta>1)$$

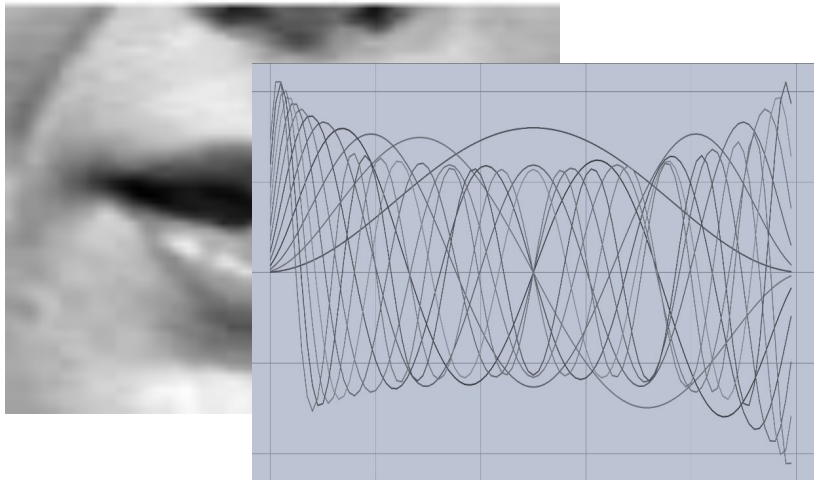
Where $(a)_k$ *is the pchhammer symbol*



Discrete orthogonal Hahn Moments

2D Hahn moments of $(n \times m)$ for an image of size $(N \times N)$ is given by *:

$$H_{nm} = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} h_m^{(\alpha,\beta)}(x, N) \cdot h_n^{(\alpha,\beta)}(y, N) \cdot f(x, y), \qquad with\, n, m = 0,1, \cdots, N-1$$

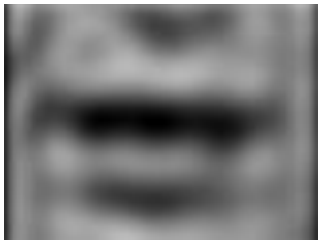


* Zhou, Jian, et al. "Image analysis by discrete orthogonal Hahn moments." *International Conference Image Analysis and Recognition*. 2005.

Image reconstruction property

The image can be constructed as follows

$$\widetilde{f(x,y)} = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1} h_m^{(\alpha,\beta)}(x,N)\cdot h_n^{(\alpha,\beta)}(y,N)\cdot H_{nm}, \qquad with\ n,m = 0,1,\cdots,N-1$$



Order 12

Order 16

Order 20

Order 32

Original: 80x60

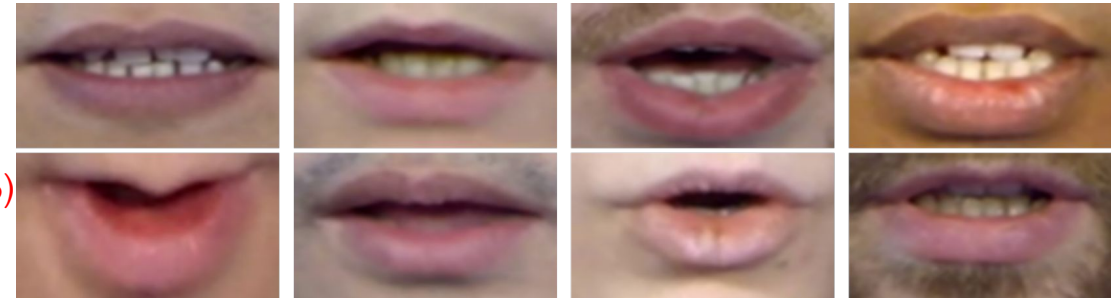
Speech Data

AVLetters (26 Alphabet letters):

780 videos for 10 speakers, every speaker utters the 26 alphabet letters three times.
20 to 40 frame per video

OuluVS2 (10 Digits sequences):

52 speakers uttering 10 digits sequences with a repetition of 3 times each.
High dimensionality (up to 250 frame per video)



Oxford-BBC Lip Reading in Wild (500 words):

500 unique words with up to 1000 utterances per word spoken by different speakers.
30 frame/video

Results

AVLetters

Order	Accuracy
32	53.41%
52	59.23%
56	55.76%
CNN Without Hahn	**39.23%**

OuluVS Digits

Order	Accuracy
32	88.72%
56	93.72%
60	92.66%
CNN Without Hahn	**42.27%**

BBC LipReading in the Wild (*order 30*)

	Top@1 Accuracy	Top@5 Accuracy	Top@10 Accuracy
HCNN (without DA)	55.86%	82.93%	89.95%
HCNN (+ flip DA)	58.02%	84.54%	90.86%