#DLUPC

Day 4 Lecture 6

Speech and Vision

http://bit.ly/dlcv2018

Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

Acknowledgments

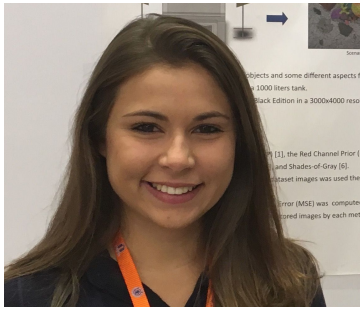
Antonio Bonafonte

Santiago Pascual

Marta R. Costa-jussà

Jose A. Rodríguez Fonollosa

Acknowledgments



Amanda Duarte

Janna Escur

Alejandro Woodward

Fran Roldan

Janna Escur, "Exploring Automatic Speech Recognition with Tensorflow" UPC ETSETB 2018.

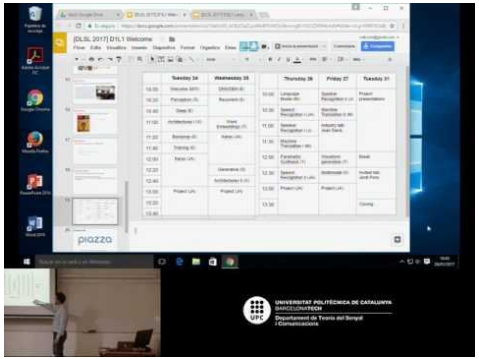
Roldan F, Pascual S, Salvador A, McGuinness K, Giro-i-Nieto X, "Speech-conditioned Face Generation with Deep Adversarial Networks" (under progress)

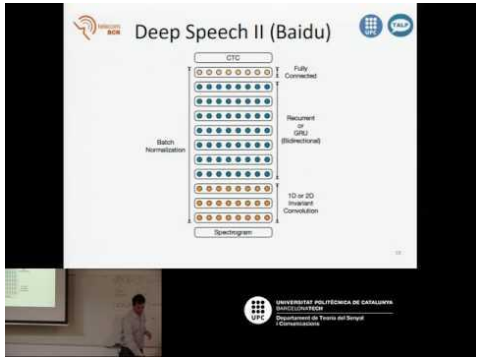
Acknowledgments

bit.ly/DLCV2018
#DLUPC

- 1st edition (2017)
- 2nd edition (2018)

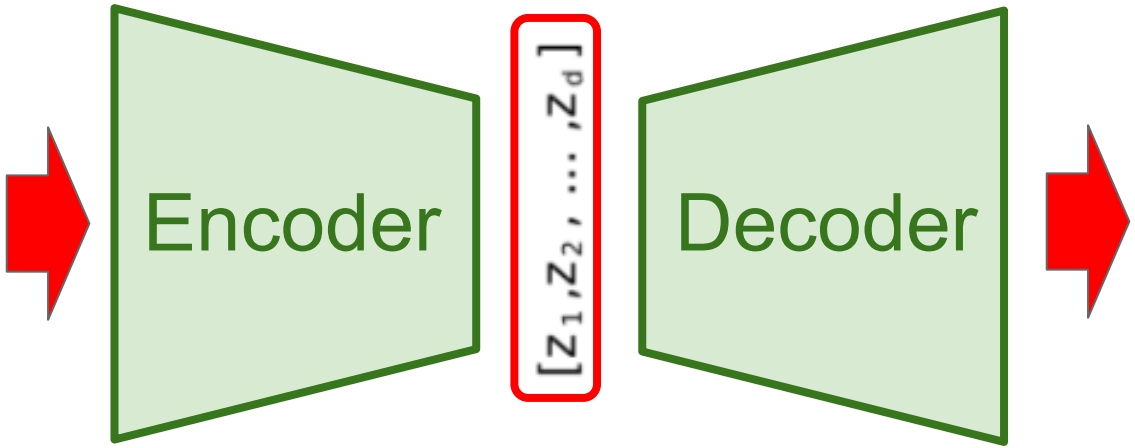
Speech Recognition



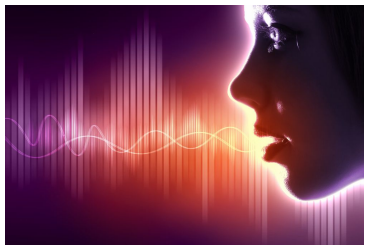


Speech Synthesis

Representation or Embedding



Representation or Embedding

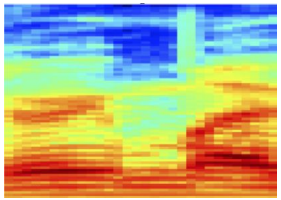


Speech

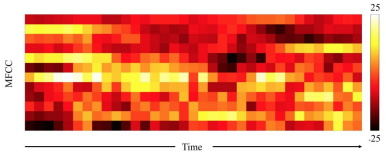
Encoder

$[z_1, z_2, \ldots, z_d]$

Raw

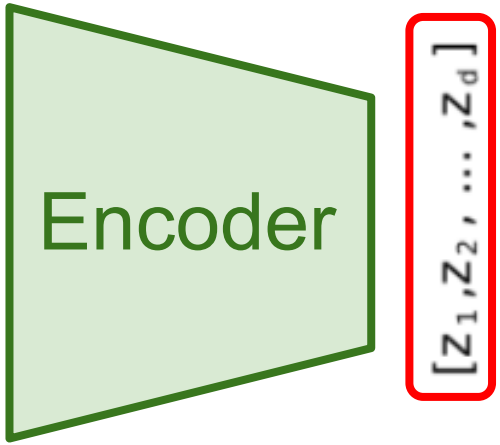


Mel spectrum



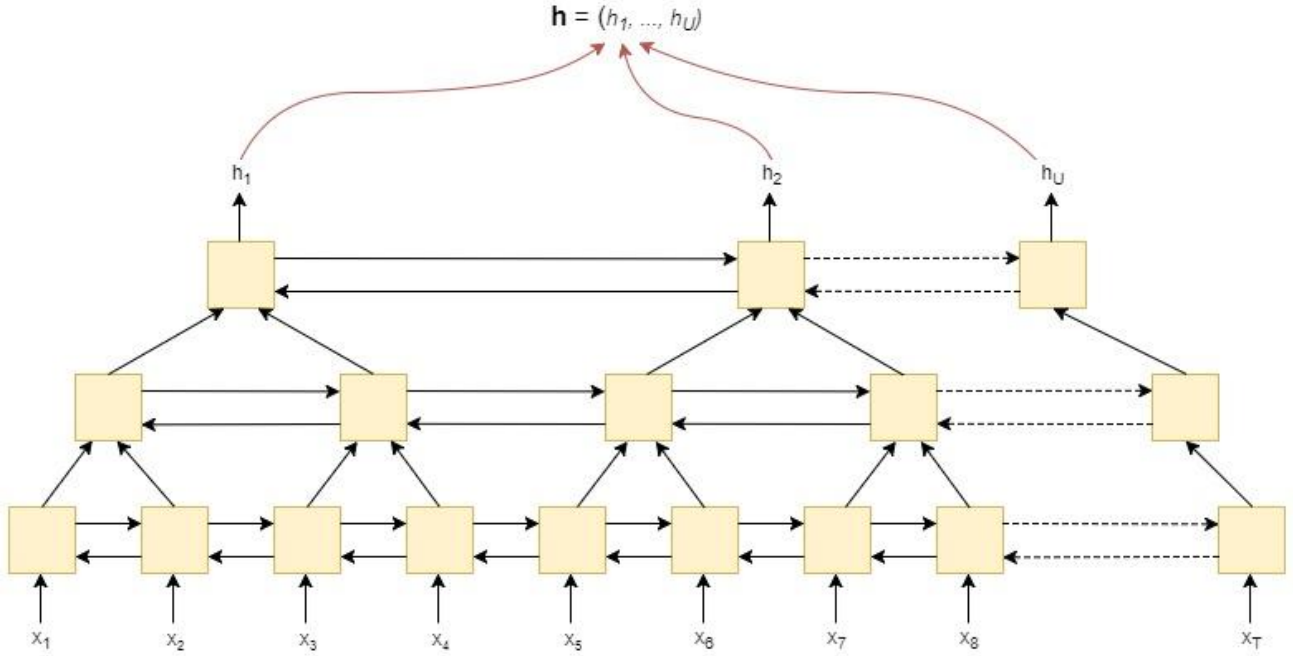
MFCC

Representation or Embedding



More details: Slides by Kishore Prahallad (CMU)

LAS Speech Encoder

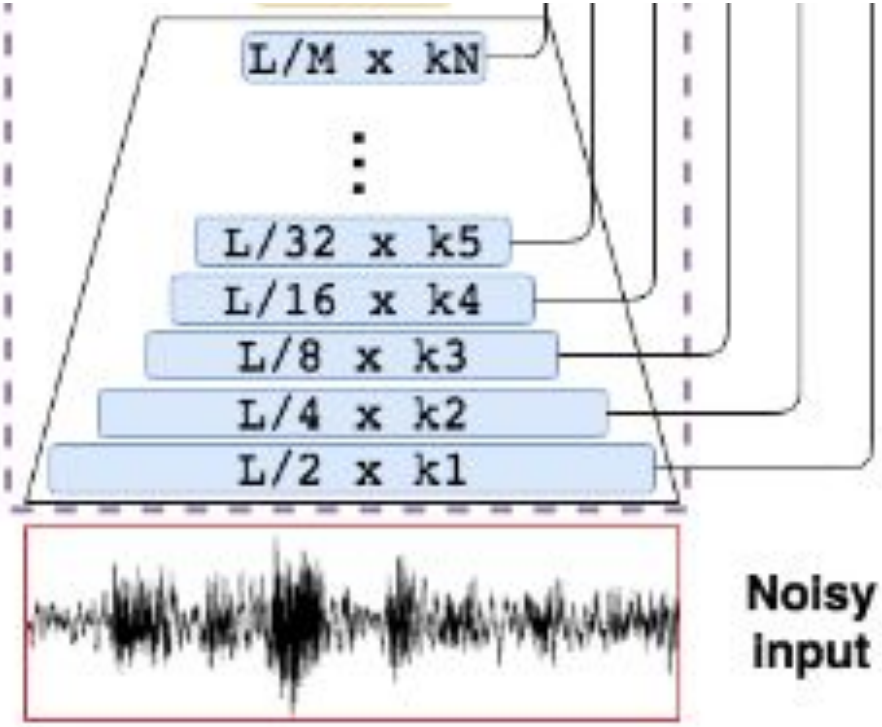


Pyramidal BLSTM

Chan, William, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition." ICASSP 2016.

Janna Escur, "Exploring Automatic Speech Recognition with Tensorflow" UPC ETSETB 2018.

SEGAN Speech Encoder



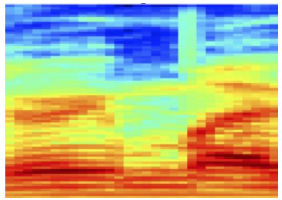
Pascual, Santiago, Antonio Bonafonte, and Joan Serra. "SEGAN: Speech enhancement generative adversarial network." Interspeech 2017.

Representation or Embedding

$[z_1, z_2, \ldots, z_d]$

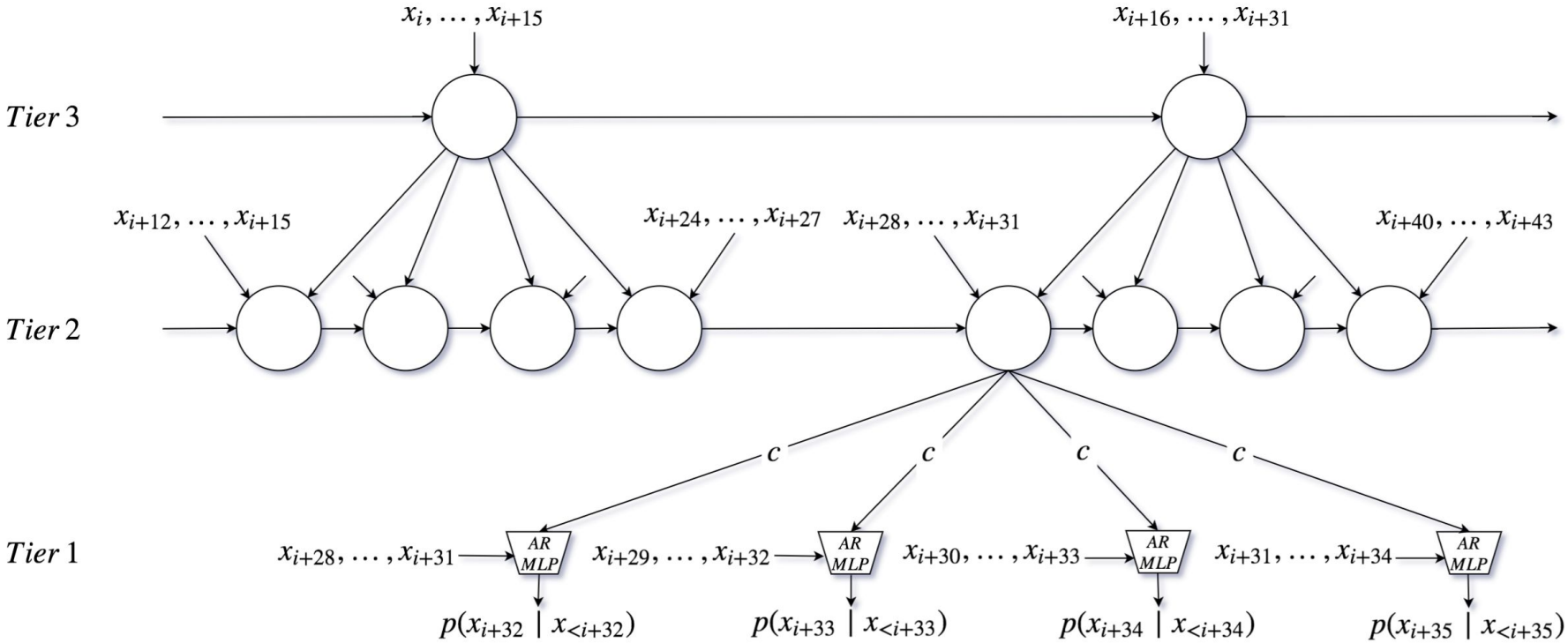
Decoder

Raw



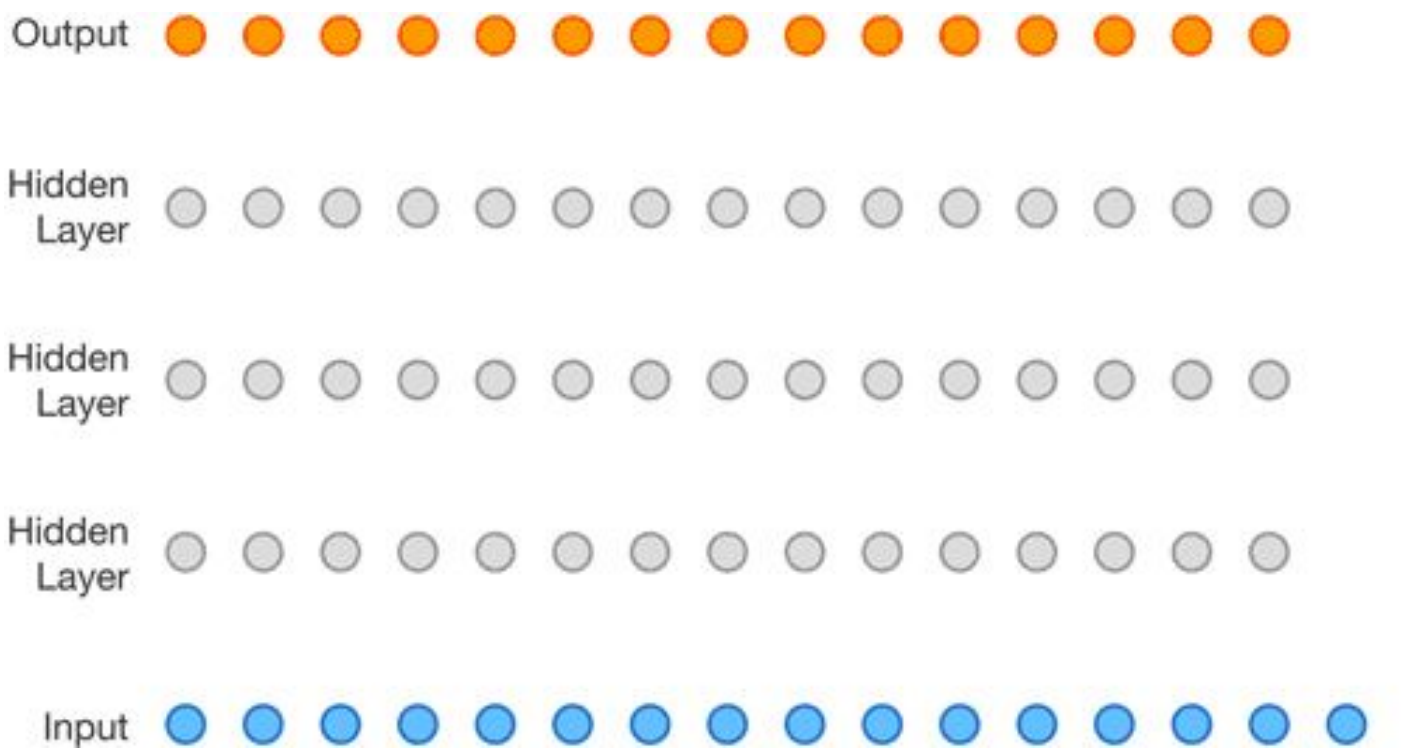
Mel spectrum

SampleRNN Speech Decoder



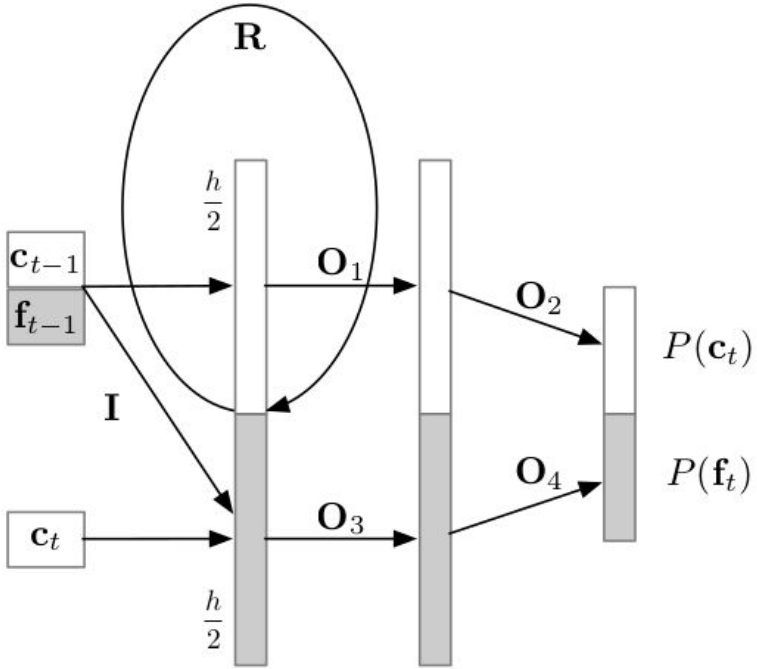
Mehri, Soroush, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. "SampleRNN: An unconditional end-to-end neural audio generation model." ICLR 2017.

WaveNet Speech Decoder



Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. "Wavenet: A generative model for raw audio." arXiv preprint arXiv:1609.03499 (2016).

WaveRNN Speech Decoder



Kalchbrenner, Nal, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu. "Efficient Neural Audio Synthesis." arXiv preprint arXiv:1802.08435 (2018).

WaveGAN Speech Decoder

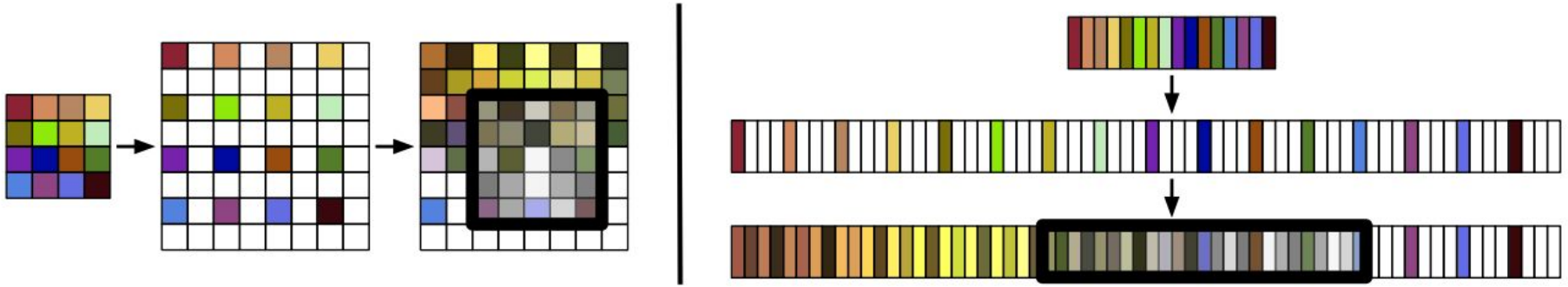
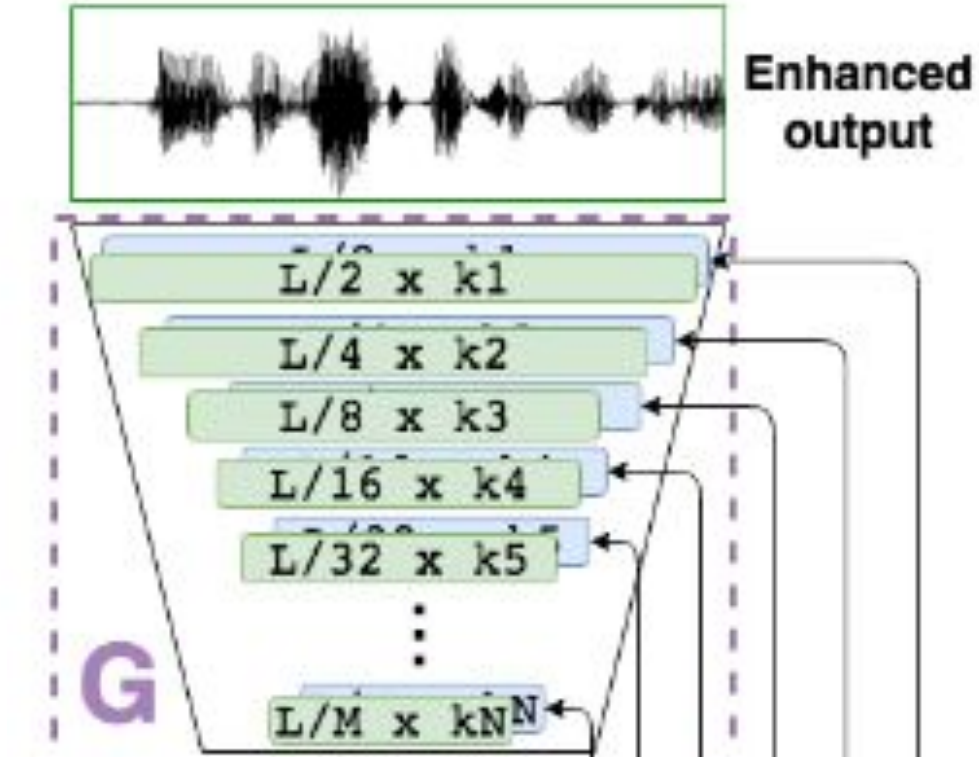


Figure 2. Depiction of the transposed convolution operation for the first layers of the DCGAN (Radford et al., 2016) (**left**) and WaveGAN (**right**) generators. DCGAN uses small (5x5), two-dimensional filters while WaveGAN uses longer (length-25), one-dimensional filters and a larger upsampling factor. The two operations have the same number of parameters and numerical operations.

Donahue, Chris, Julian McAuley, and Miller Puckette. "Synthesizing Audio with Generative Adversarial Networks." arXiv preprint arXiv:1802.04208 (2018).

SEGAN Speech Decoder



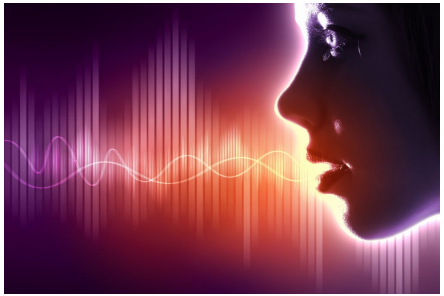
Pascual, Santiago, Antonio Bonafonte, and Joan Serra. "SEGAN: Speech enhancement generative adversarial network." Interspeech 2017.

Representation or Embedding

Encoder

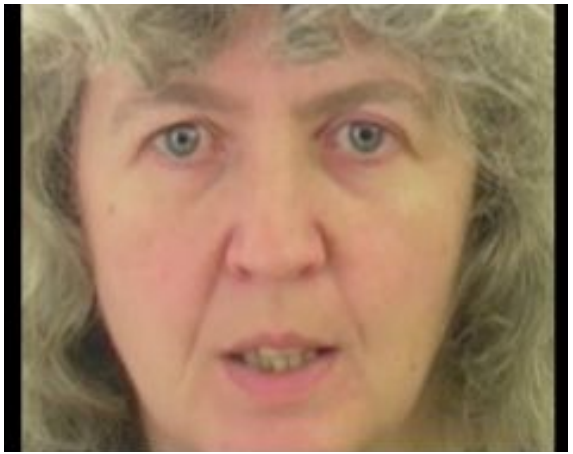
$[z_1, z_2, \ldots, z_d]$

Decoder

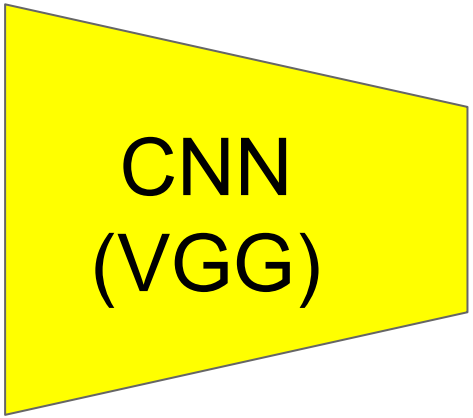


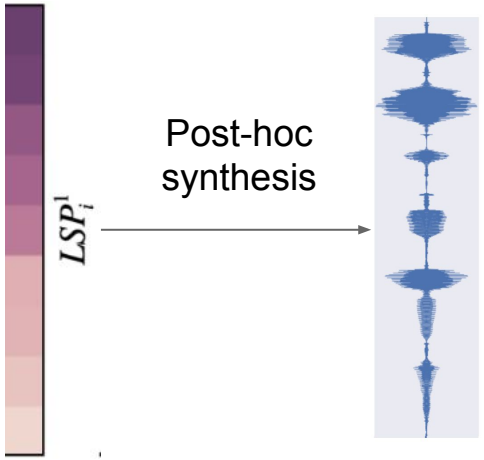
Ephrat, Ariel, Tavi Halperin, and Shmuel Peleg. "Improved speech reconstruction from silent video." In ICCV 2017 Workshop on Computer Vision for Audio-Visual Media. 2017.

Speech Reconstruction from Video



Frame from a silent video

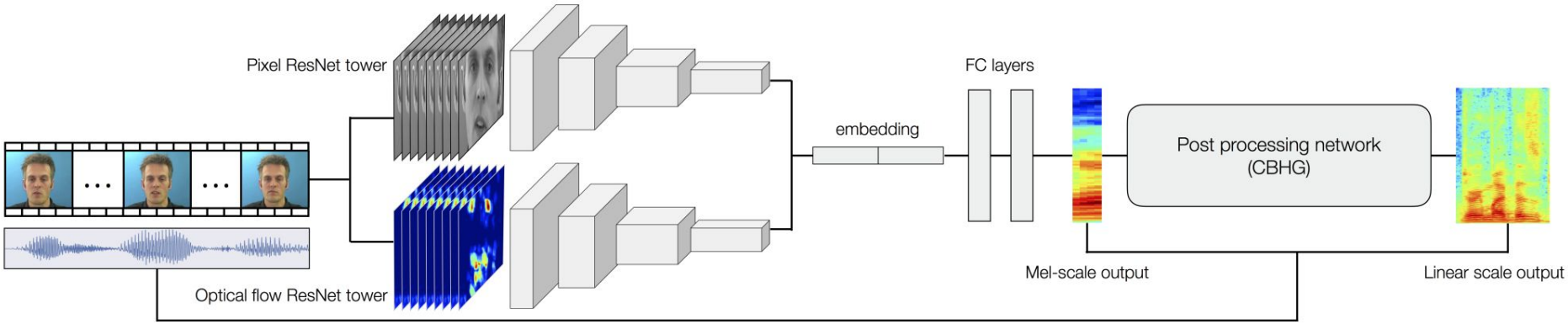




Audio feature

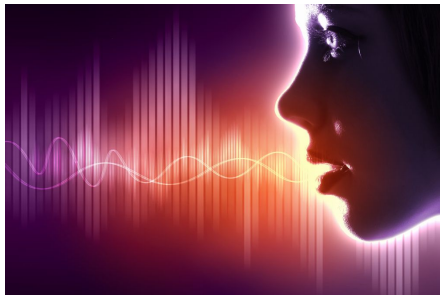
Ephrat et al. Vid2speech: Speech Reconstruction from Silent Video. *ICASSP 2017*

Speech Reconstruction from Video



Ephrat, Ariel, Tavi Halperin, and Shmuel Peleg. "Improved speech reconstruction from silent video." In ICCV 2017 Workshop on Computer Vision for Audio-Visual Media. 2017.

Representation or Embedding

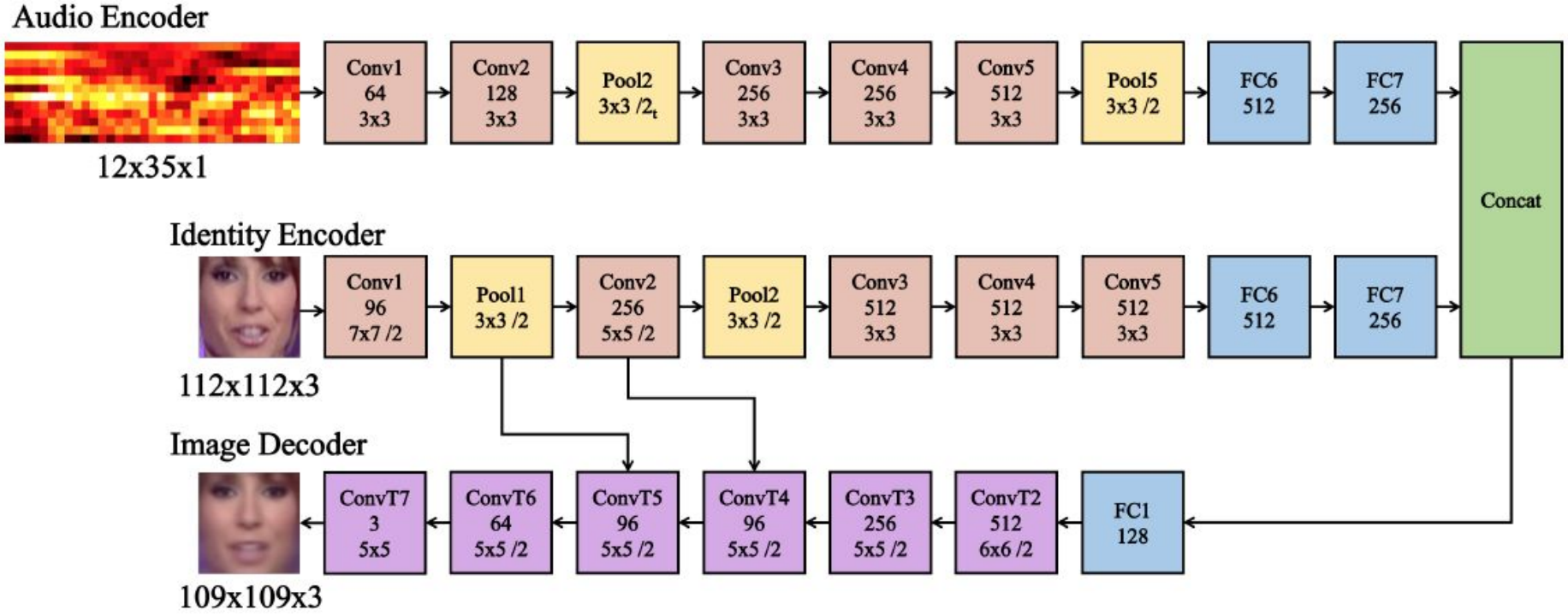


Encoder

$[z_1, z_2, \ldots, z_d]$

Decoder

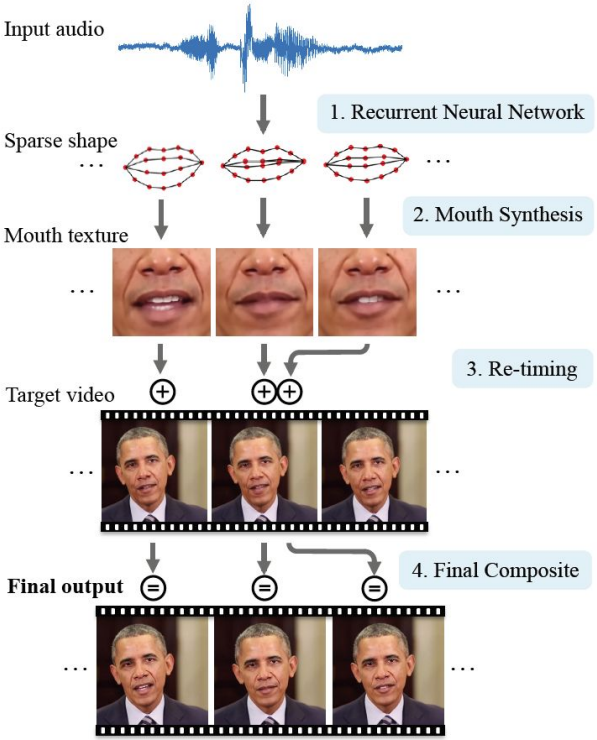
Speech to Frame Synthesis (face pixels)



Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.

Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.

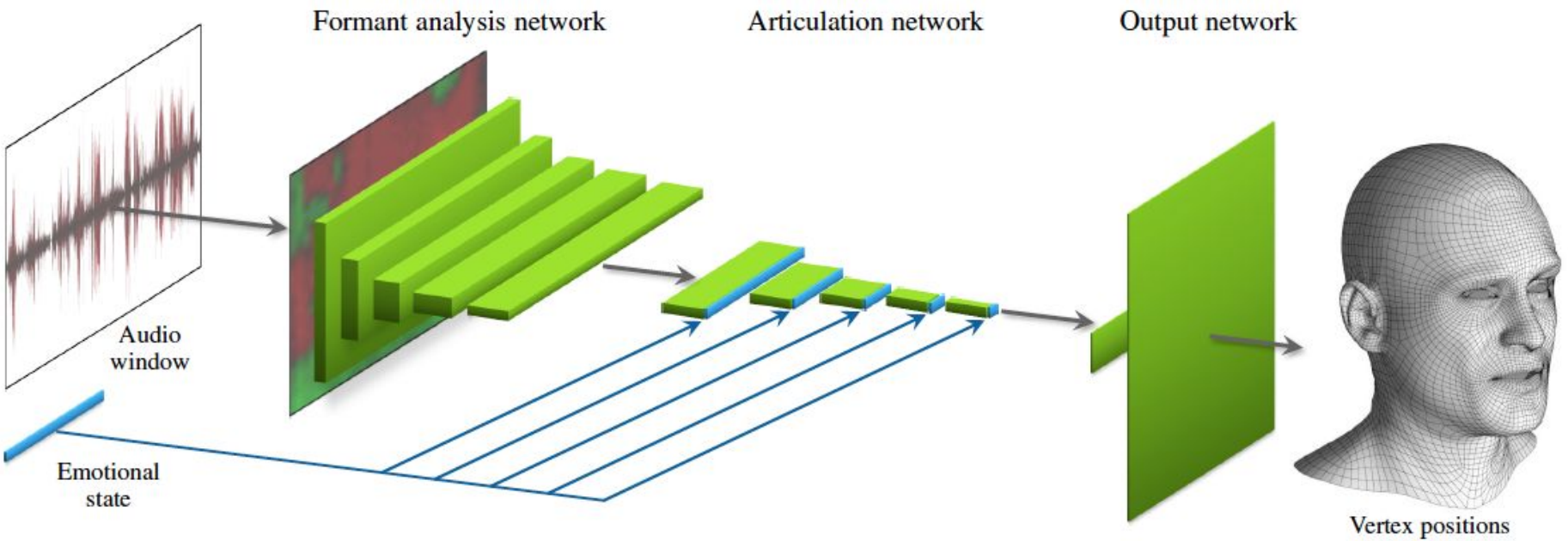
Speech to Video Synthesis (lip keypoints)



Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing Obama: learning lip sync from audio." SIGGRAPH 2017.

Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

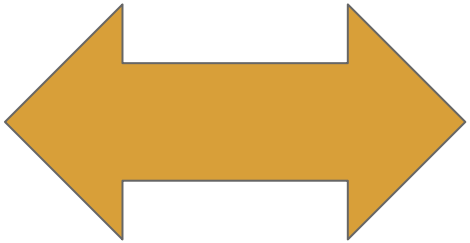
Speech to Video Synthesis (vertex positions)

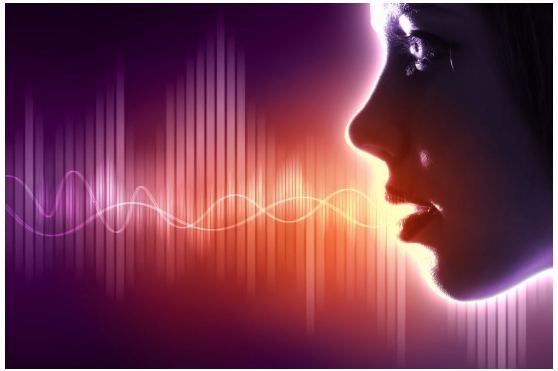


Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

Karras, Tero, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." SIGGRAPH 2017

Vision



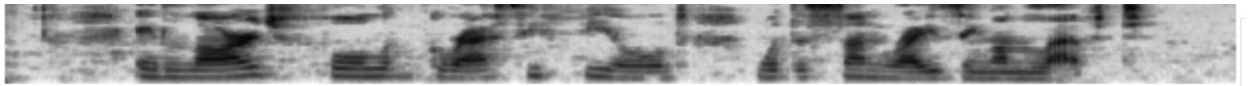


Speech

Matching speech to images

Humans understand speech much earlier than text, could computers do the same ?

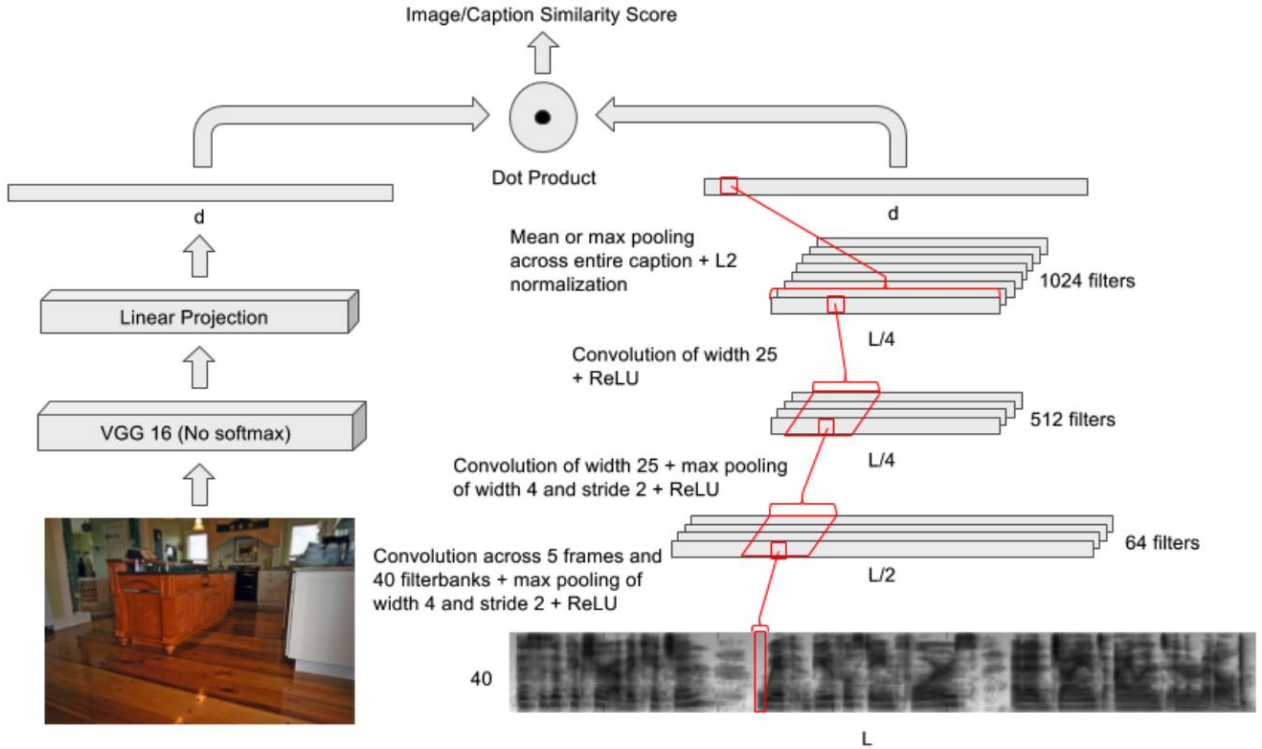
Large dataset (120,000) of speech description of images from Places dataset.



Harwath, David, Antonio Torralba, and James Glass. "Unsupervised learning of spoken language with visual context." NIPS 2016. [talk]

Matching speech to images

Train a visual & speech networks with pairs of (non-)corresponding images & speech.

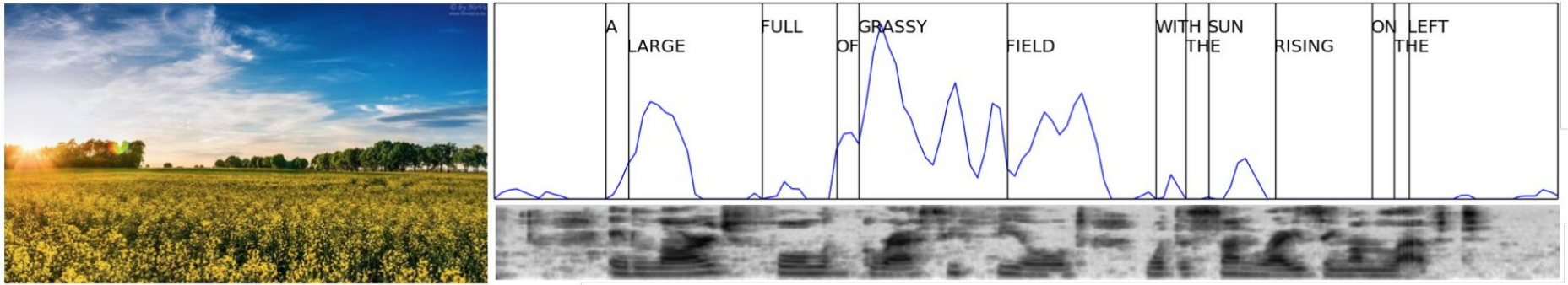


Harwath, David, Antonio Torralba, and James Glass. "Unsupervised learning of spoken language with visual context." NIPS 2016. [talk]

Matching speech to images

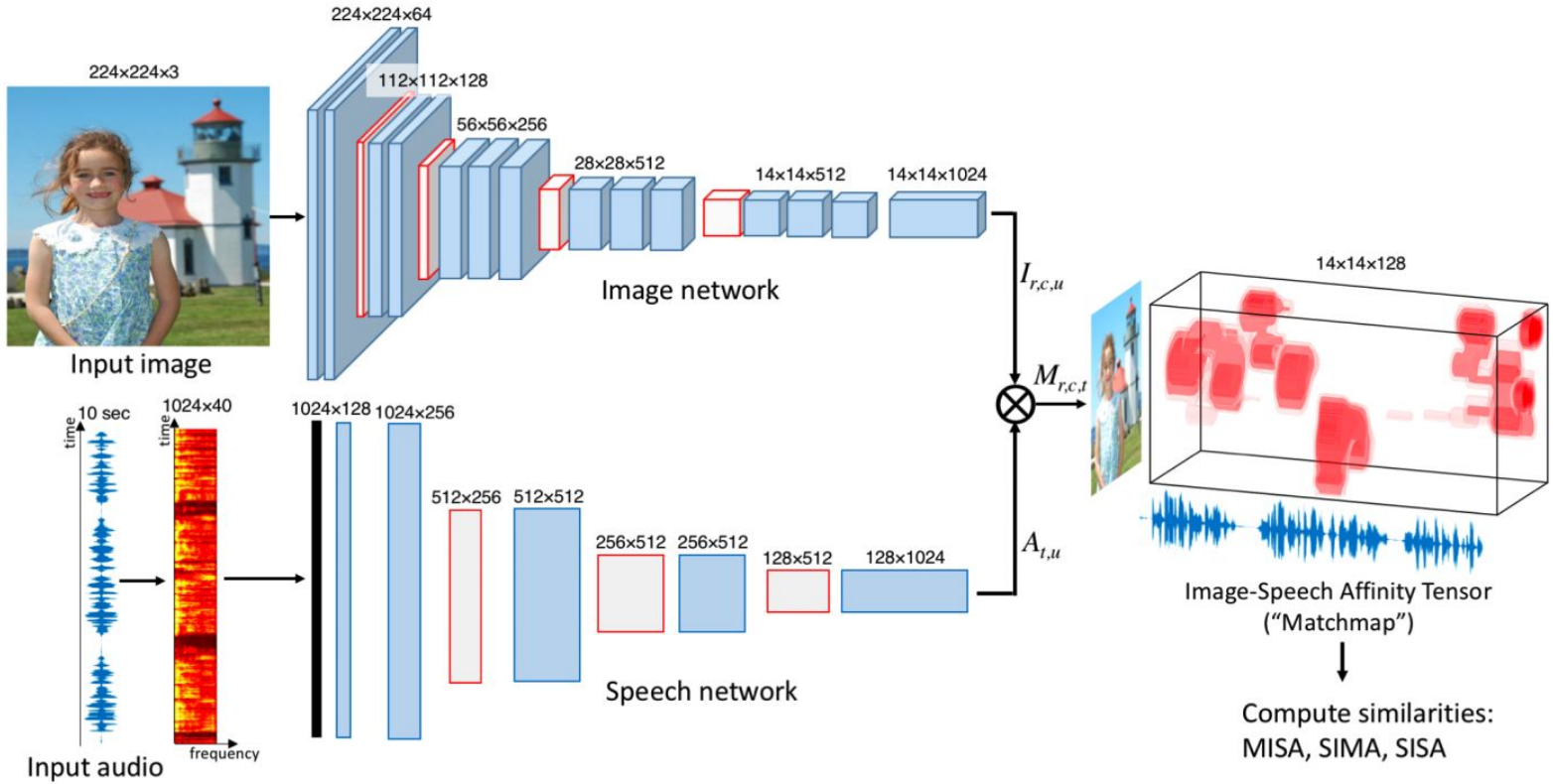
Similarity curve show which regions of the spectrogram are relevant for the image.

Important: no text transcriptions used during the training !!



Harwath, David, Antonio Torralba, and James Glass. "Unsupervised learning of spoken language with visual context." NIPS 2016. [talk]

Matching speech to objects (heatmap)



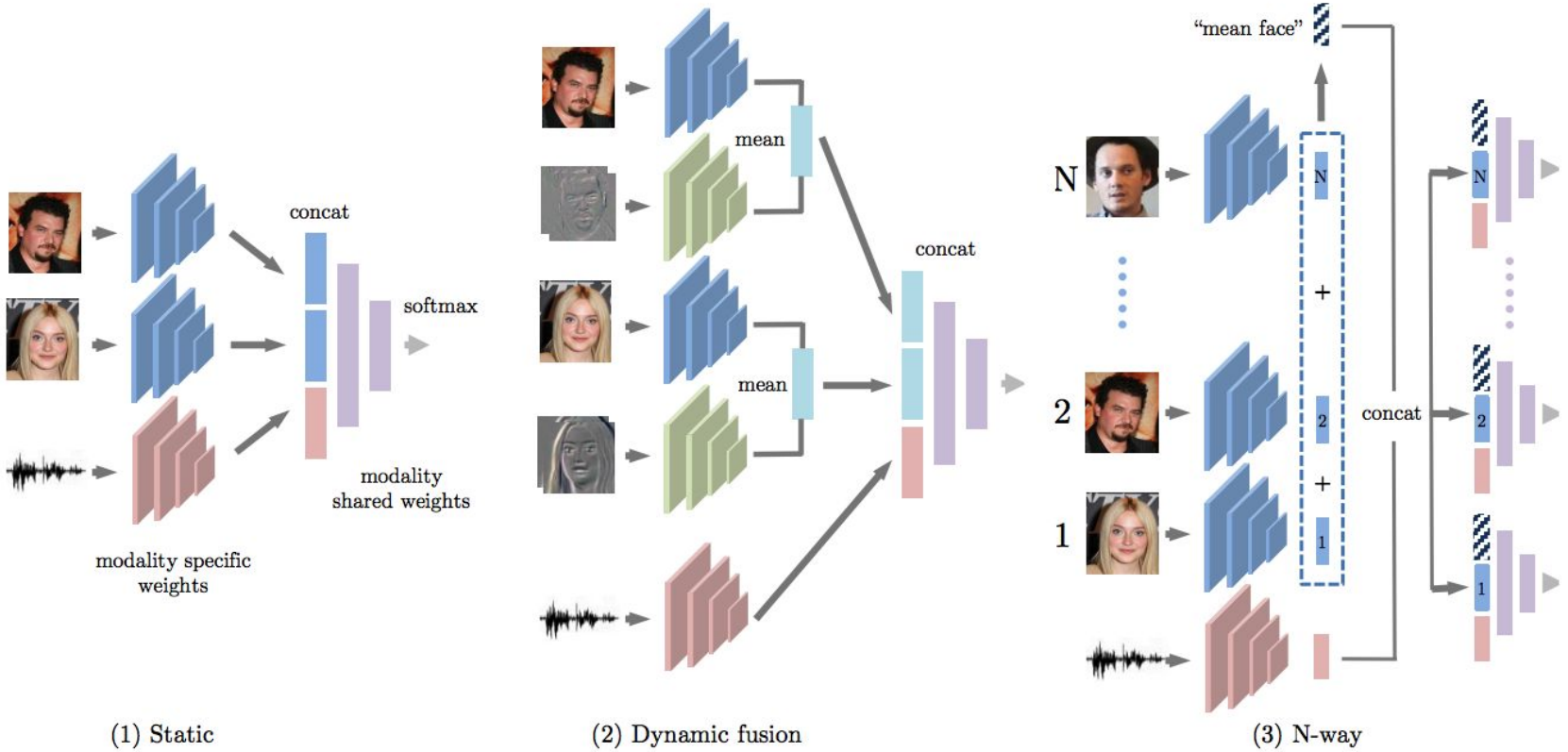
Harwath, David, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input." arXiv preprint arXiv:1804.01452 (2018).

Matching speech to objects (heatmap)

Regions matching the spoken word “WOMAN”:

Harwath, David, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. "Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input." arXiv preprint arXiv:1804.01452 (2018).

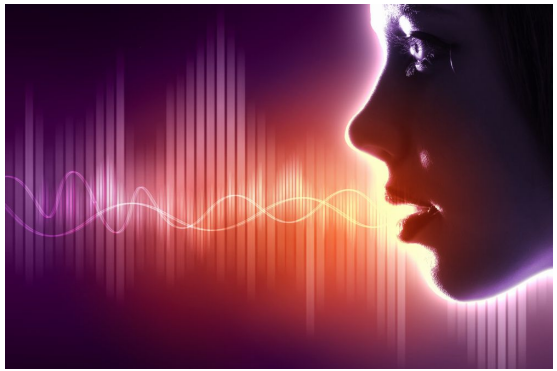
Matching speech to objects (faces)



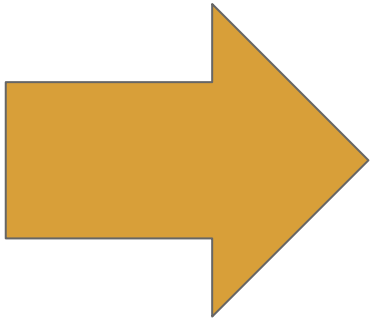
Nagrani, Arsha, Samuel Albanie, and Andrew Zisserman. "Seeing Voices and Hearing Faces: Cross-modal biometric matching." CVPR 2018. [video]

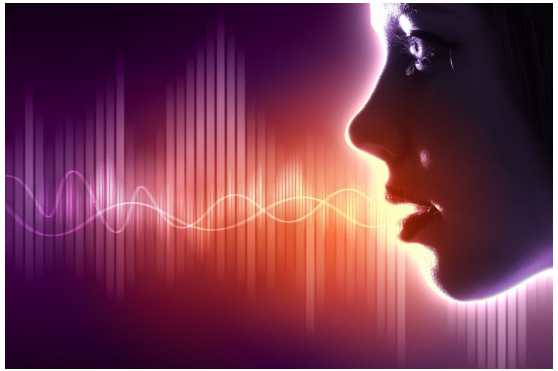
Nagrani, Arsha, Samuel Albanie, and Andrew Zisserman. "Seeing Voices and Hearing Faces: Cross-modal biometric matching." CVPR 2018. [video]

Vision



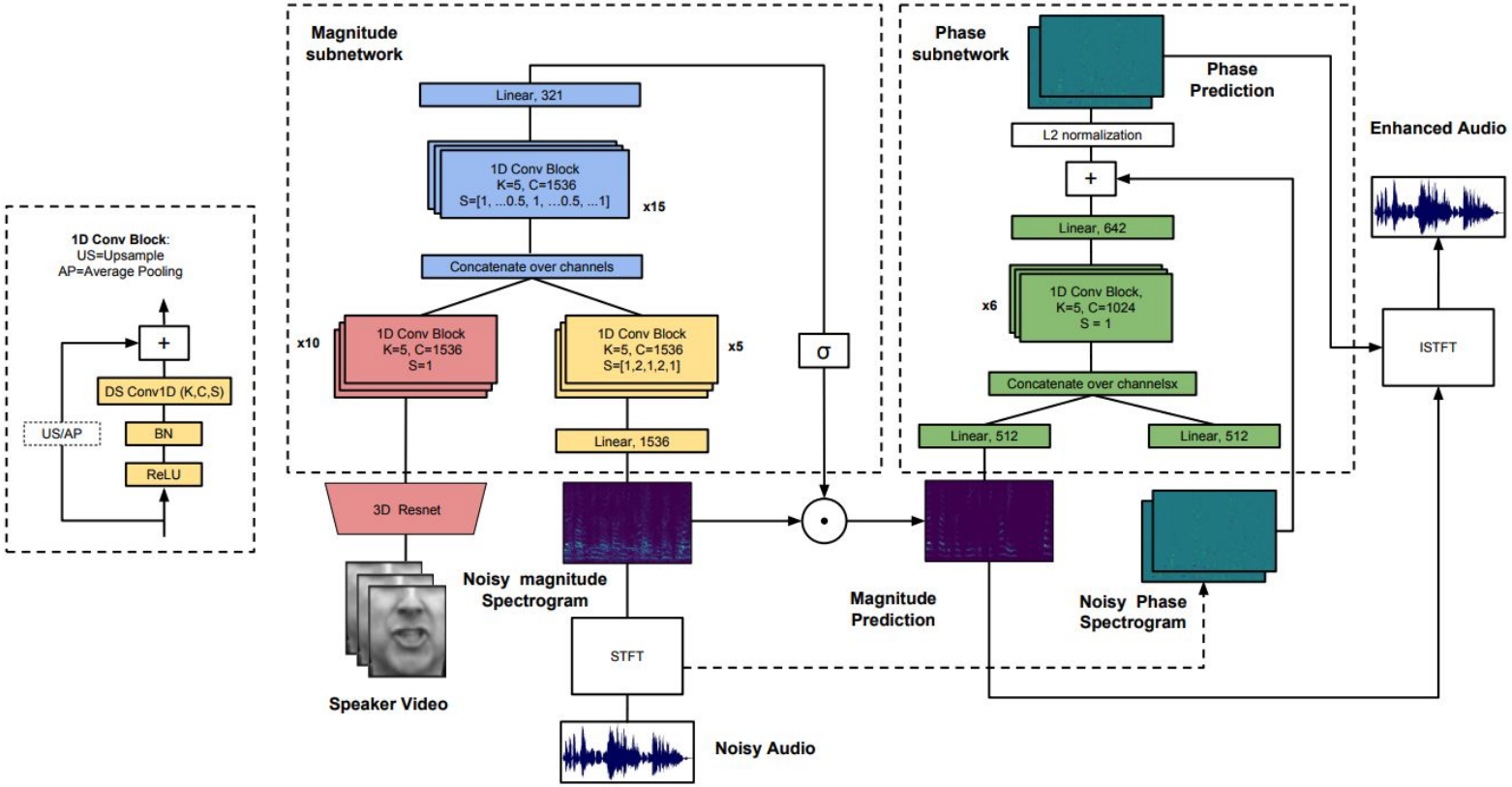
Speech





Speech

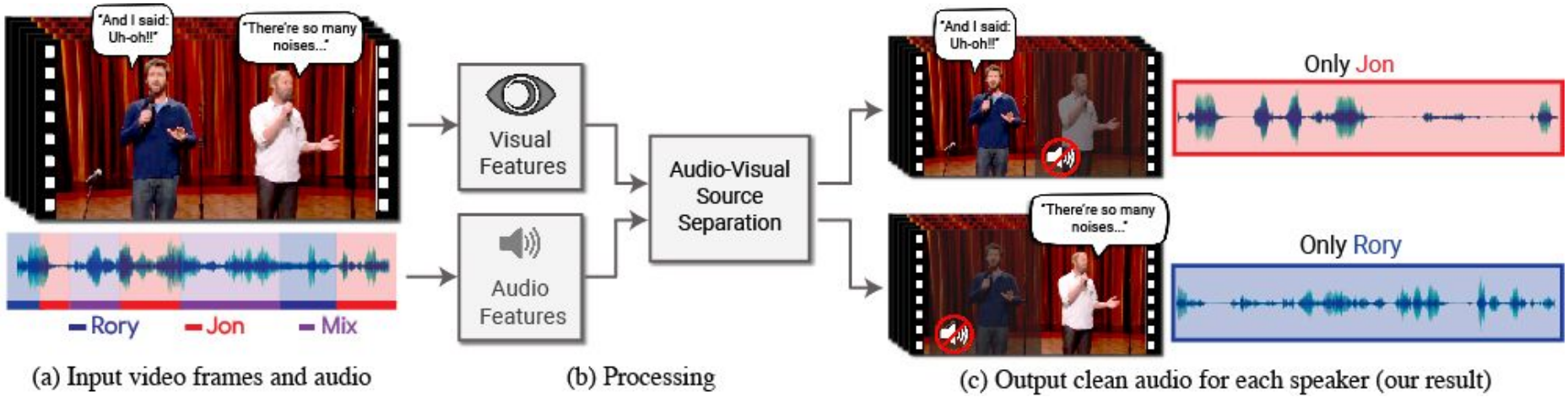
Speech Separation with Vision (lips)



Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "The Conversation: Deep Audio-Visual Speech Enhancement." Interspeech 2018.

Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "The Conversation: Deep Audio-Visual Speech Enhancement." Interspeech 2018..

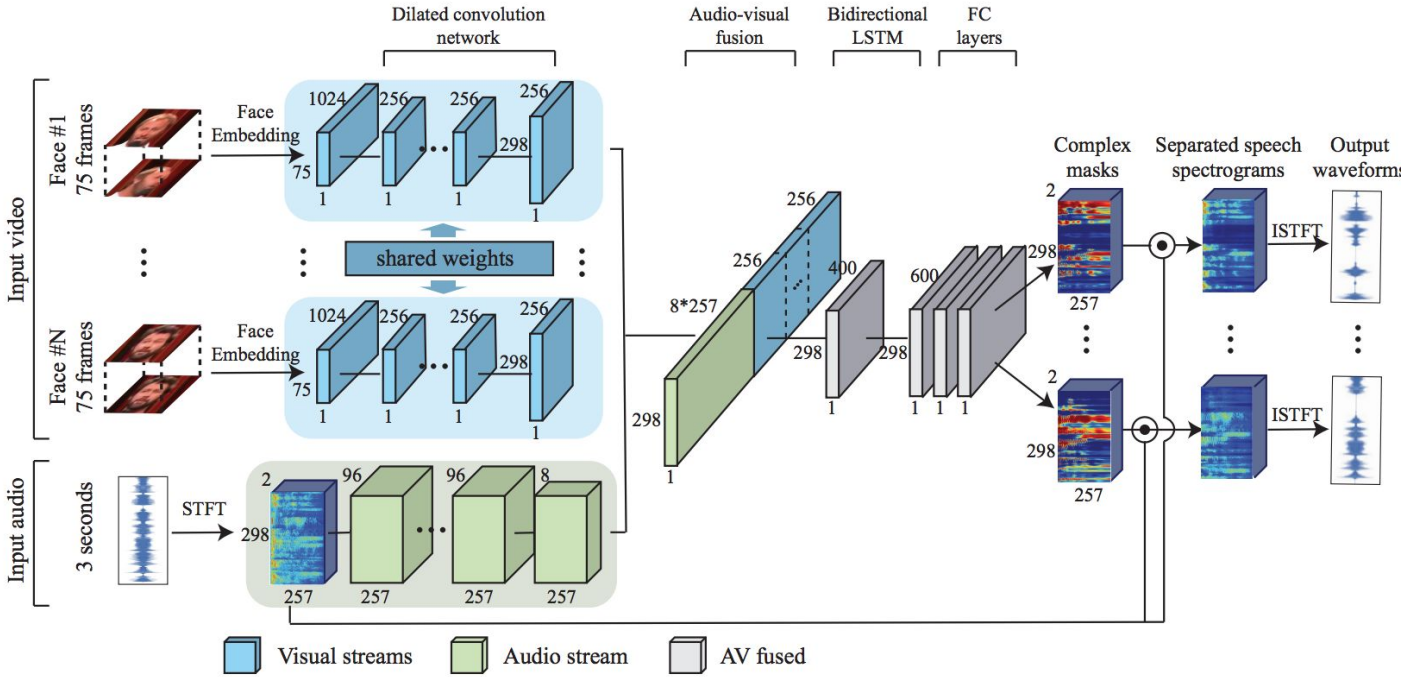
Speech Separation with Vision (faces)



(a) Input video frames and audio (b) Processing (c) Output clean audio for each speaker (our result)

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman and Michael Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation" SIGGRAPH 2018.

Speech Separation with Vision (faces)



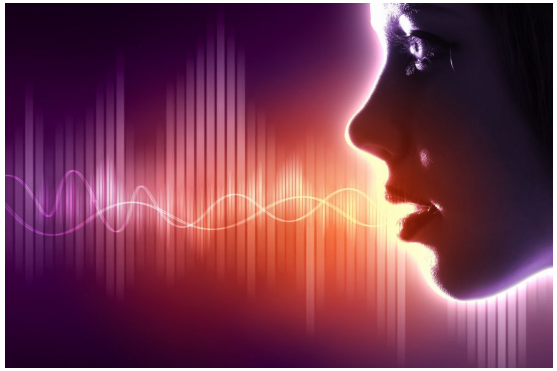
Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman and Michael Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation" SIGGRAPH 2018.

Video source: Team Coco, https://www.youtube.com/watch?v=UT7h4nRcWjU

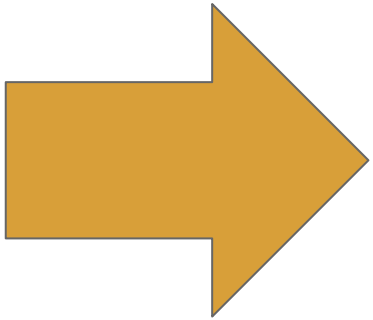
Audio-Visual Speech Separation Results | Comparison with Audio-Only | Comparison with Audio-Visual Methods | Application to Video Transcription

Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman and Michael Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation" SIGGRAPH 2018.

Vision

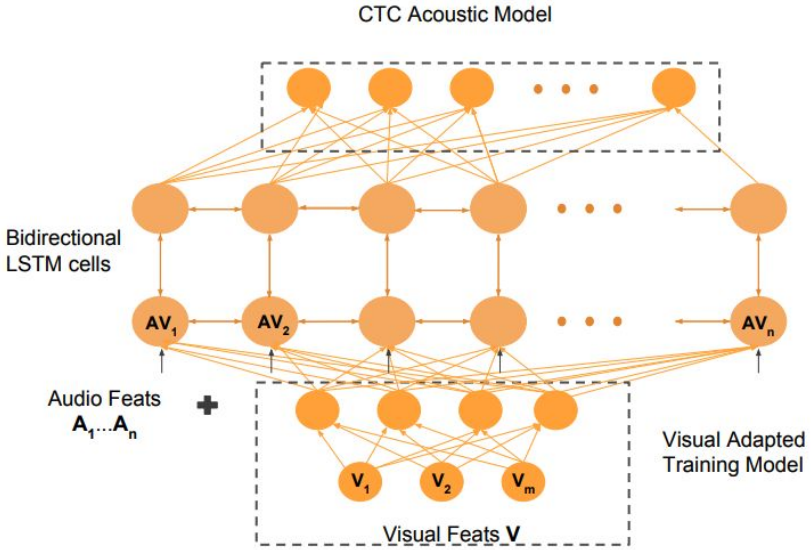


Speech

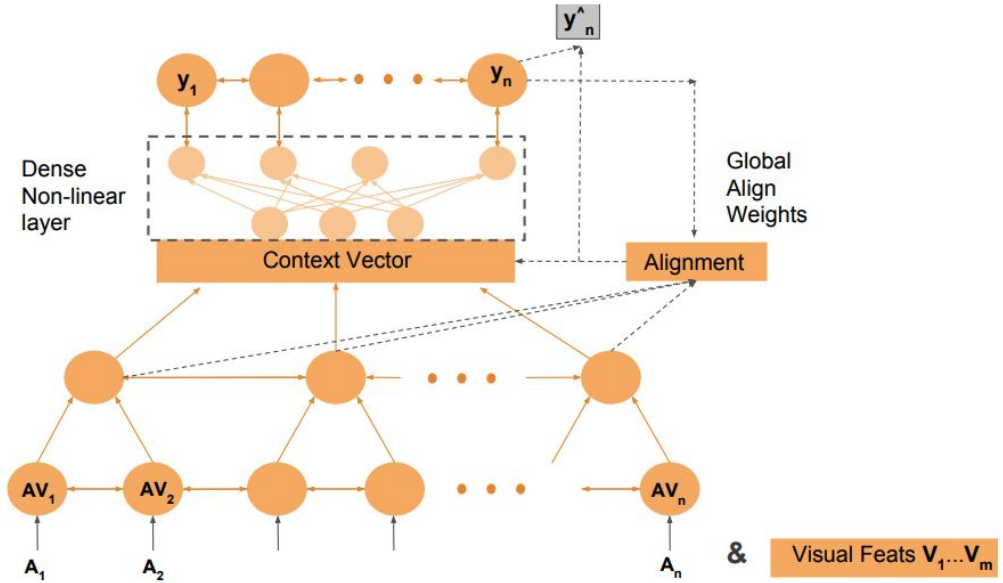


Text

Speech Recognition with vision

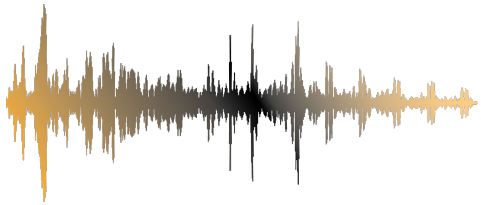


(a) CTC model architecture with adaptation



(b) S2S model architecture with global attention and adaptation

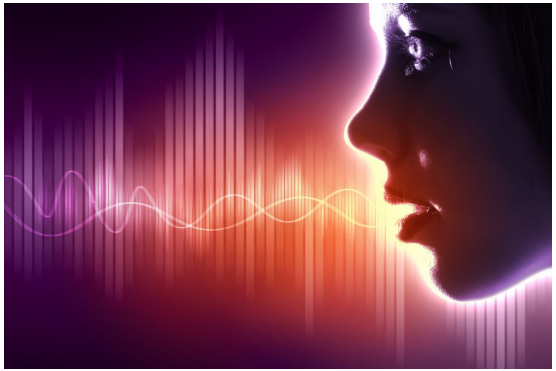
Palaskar, Shruti, Ramon Sanabria, and Florian Metze. "End-to-End Multimodal Speech Recognition." Interspeech 2018.



Audio

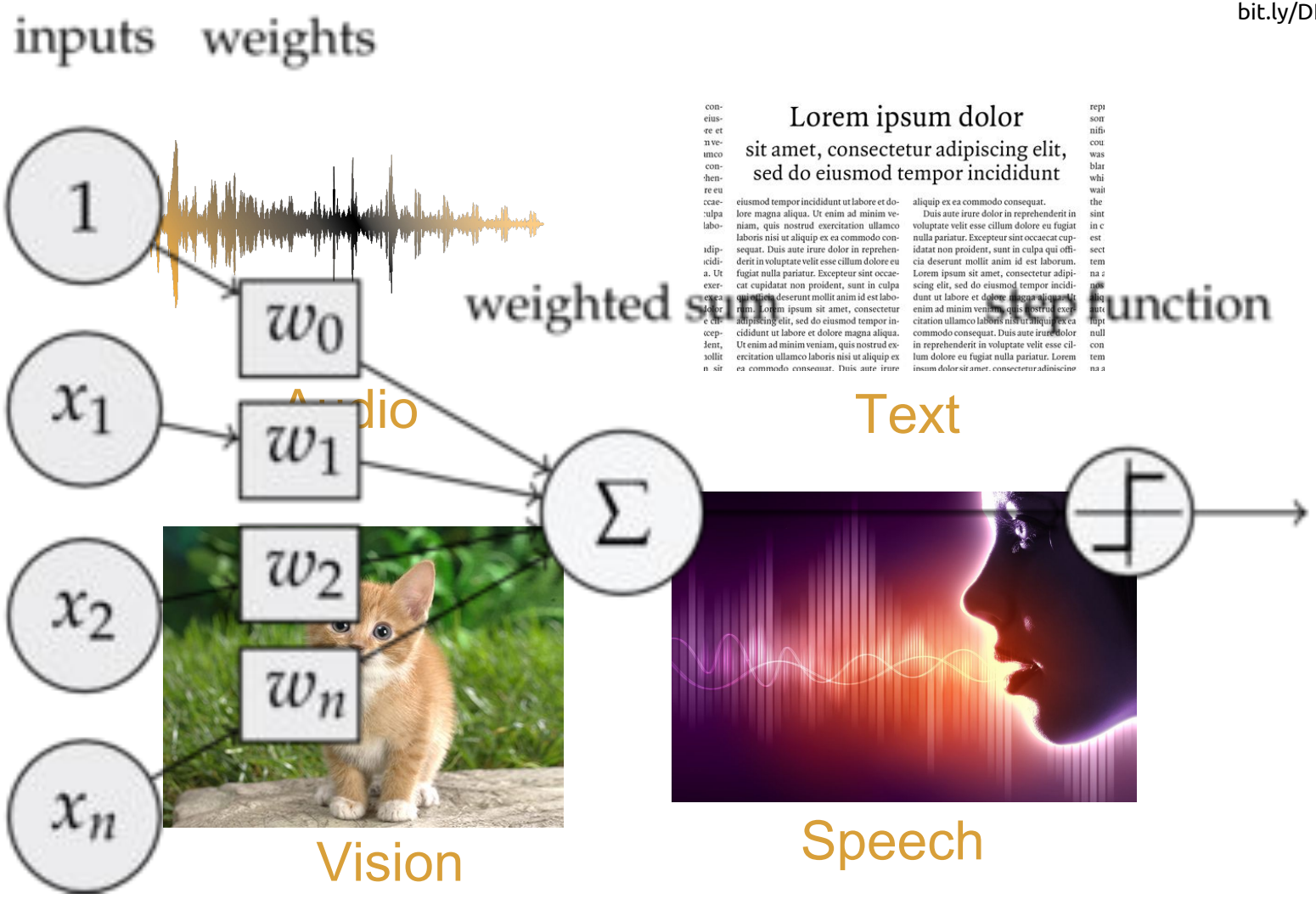
Text

Vision

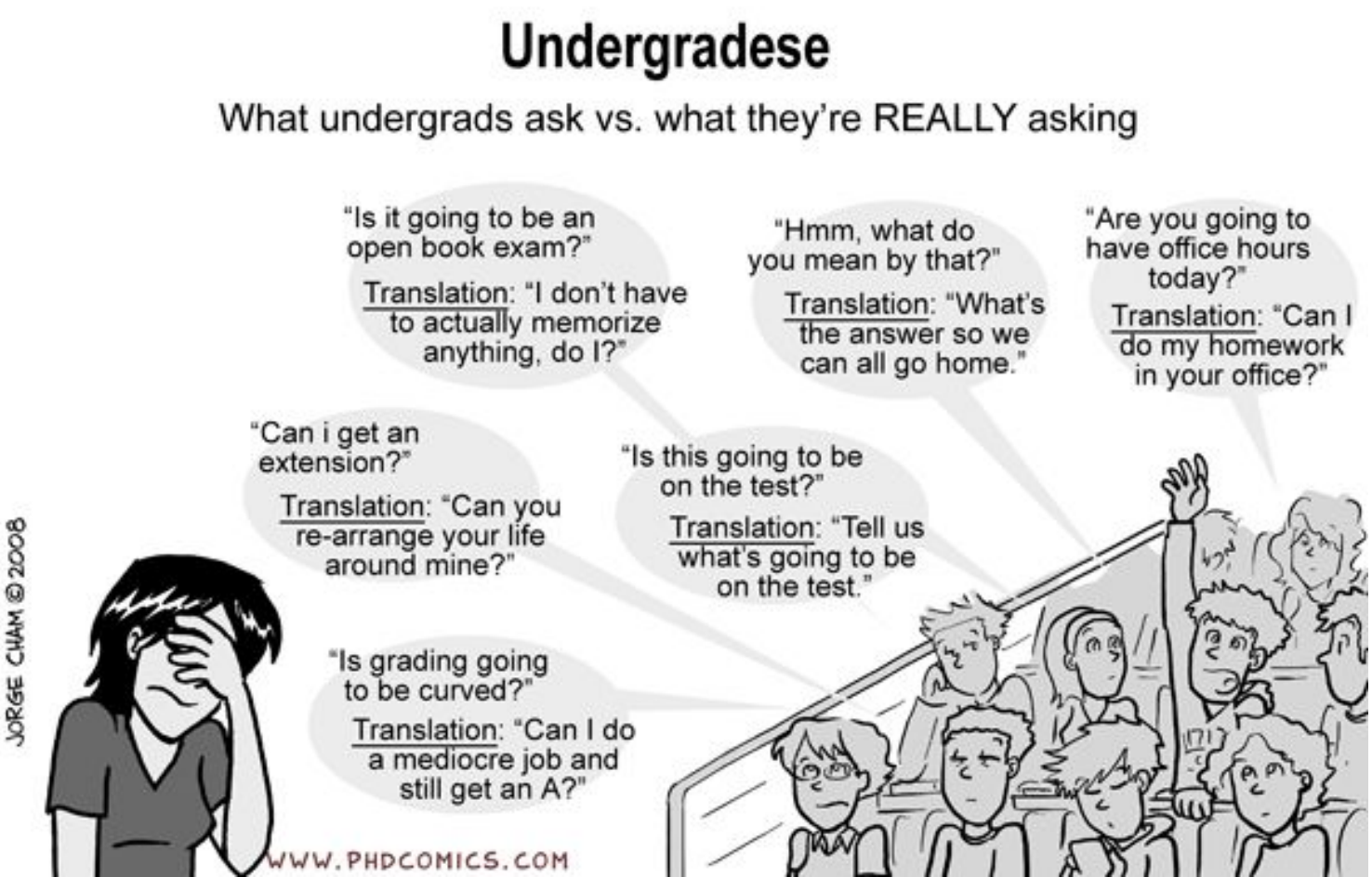


Speech


dolor
Vision
Speech


inputs
weights
1
x_1
x_2
x_n
w_0
w_1
w_2
w_n
Σ
Audio
Lorem ipsum dolor
sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt
Text
Vision
Speech

Questions


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM