Lecture 15: Multimodal learning

Announcements

- Project proposal due after spring break
- Midterm course evaluation due tomorrow

Coffee Shop

Chair

Table

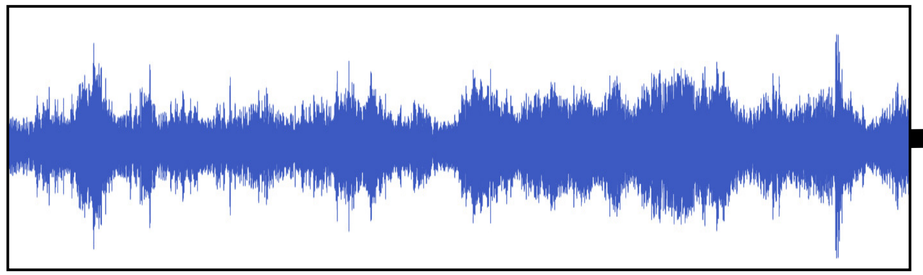

SHIPPING
RECEIVING

Cue combination

Vision

1
1
0

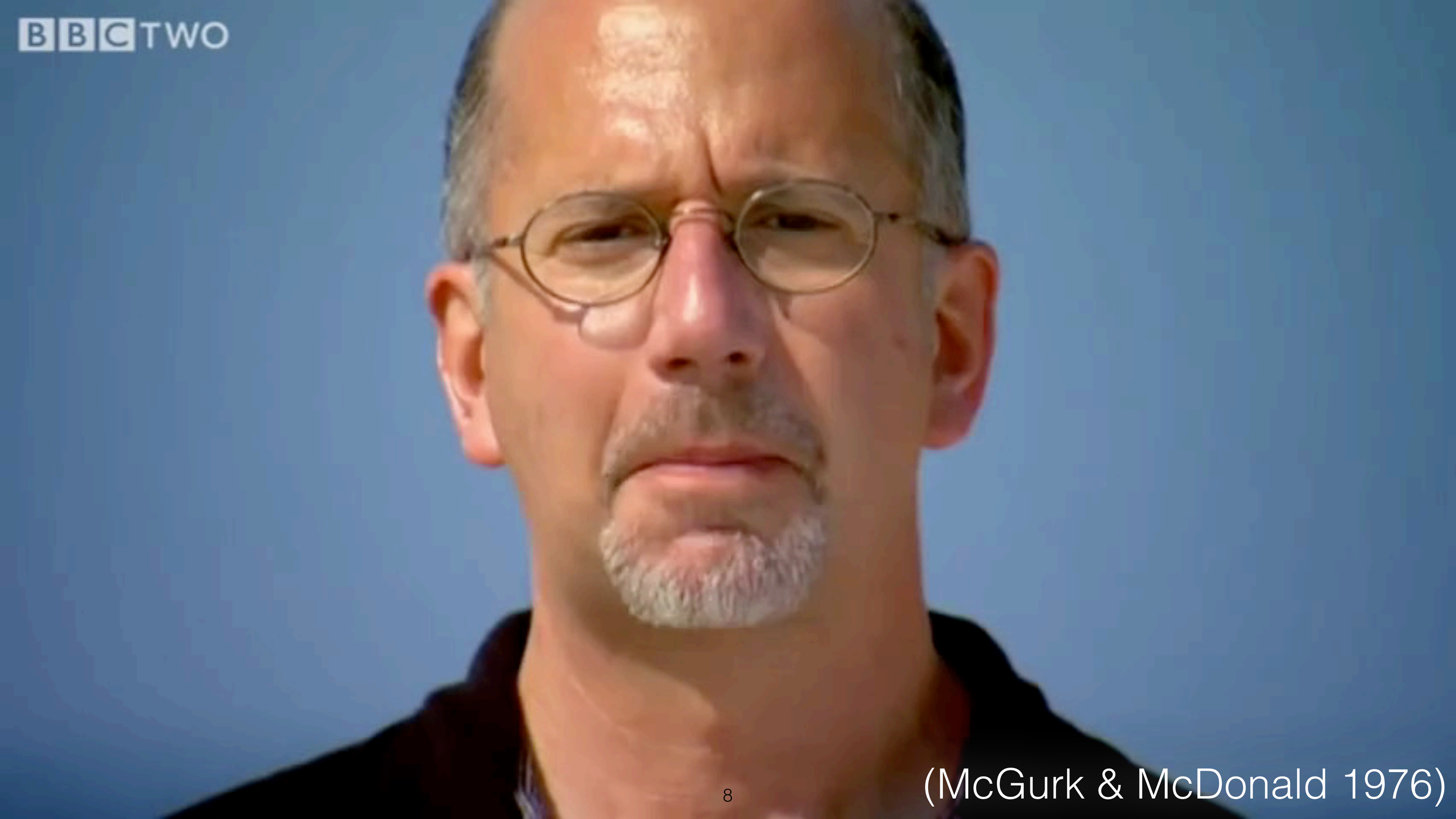
+



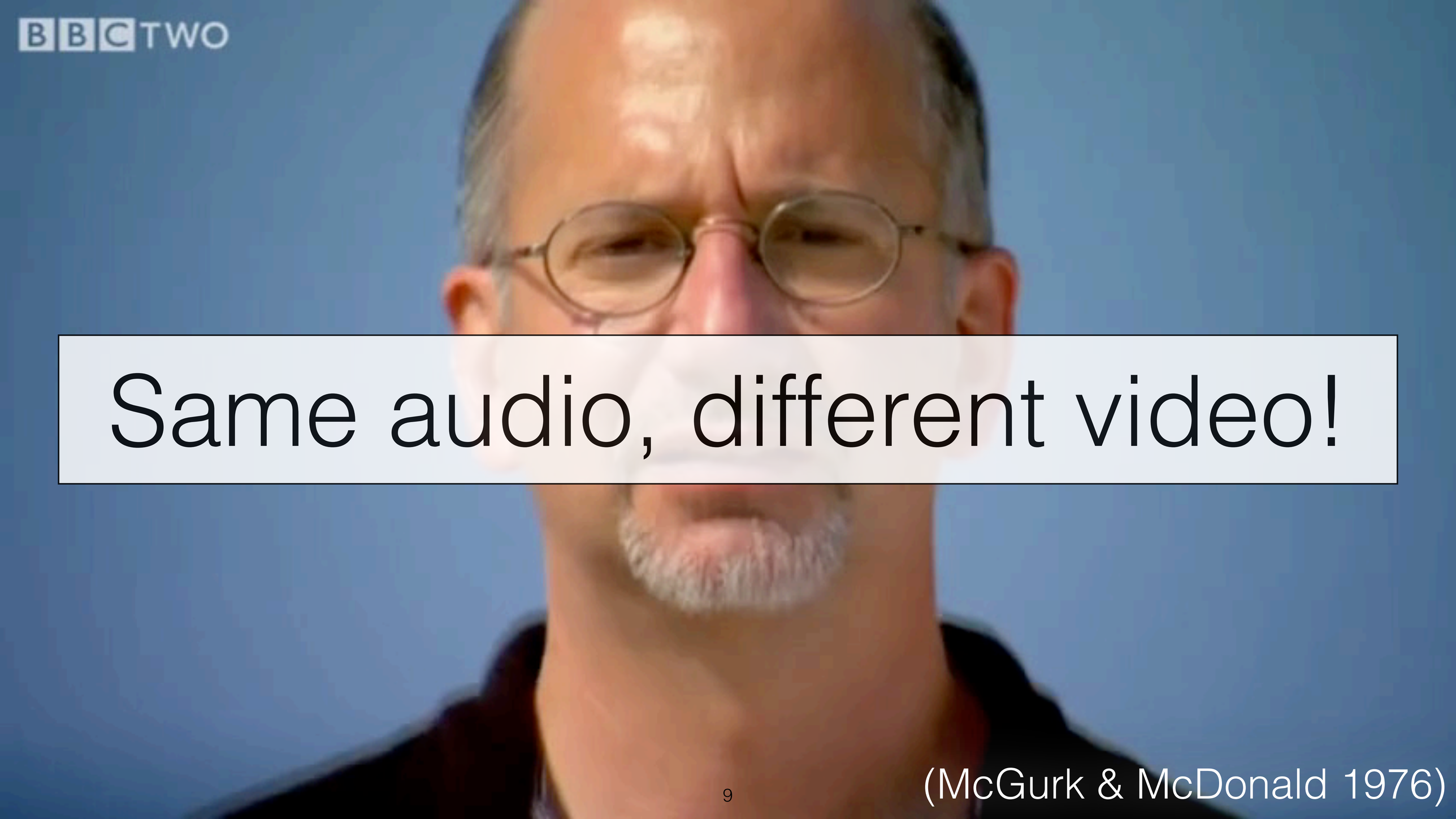
Hearing

0
1
0

(Yuhas et al. 1989)



(McGurk & McDonald 1976)



Same audio, different video!

(McGurk & McDonald 1976)

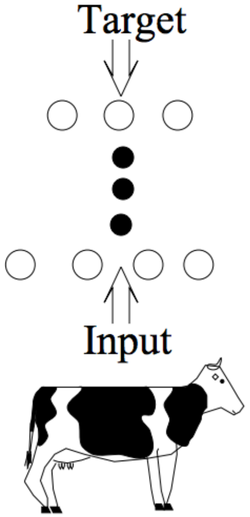
(McGurk & McDonald 1976)

Multisensory self-supervision

Supervised

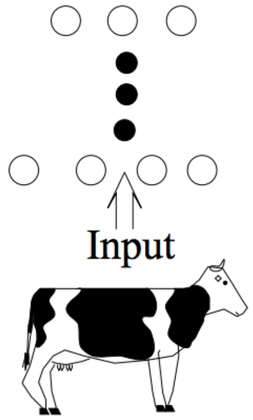
- implausible label

"COW"



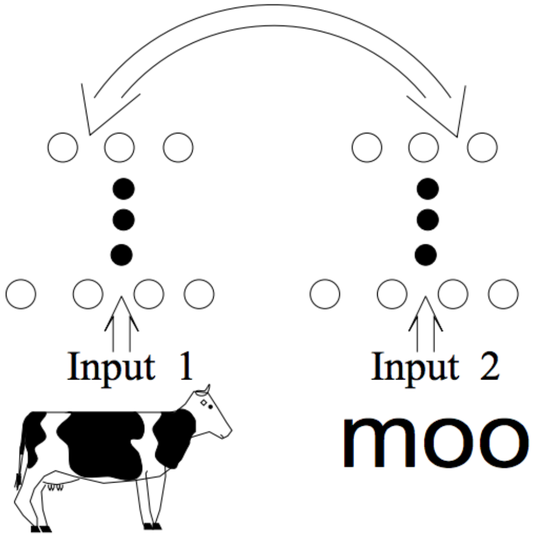
Unsupervised

- limited power



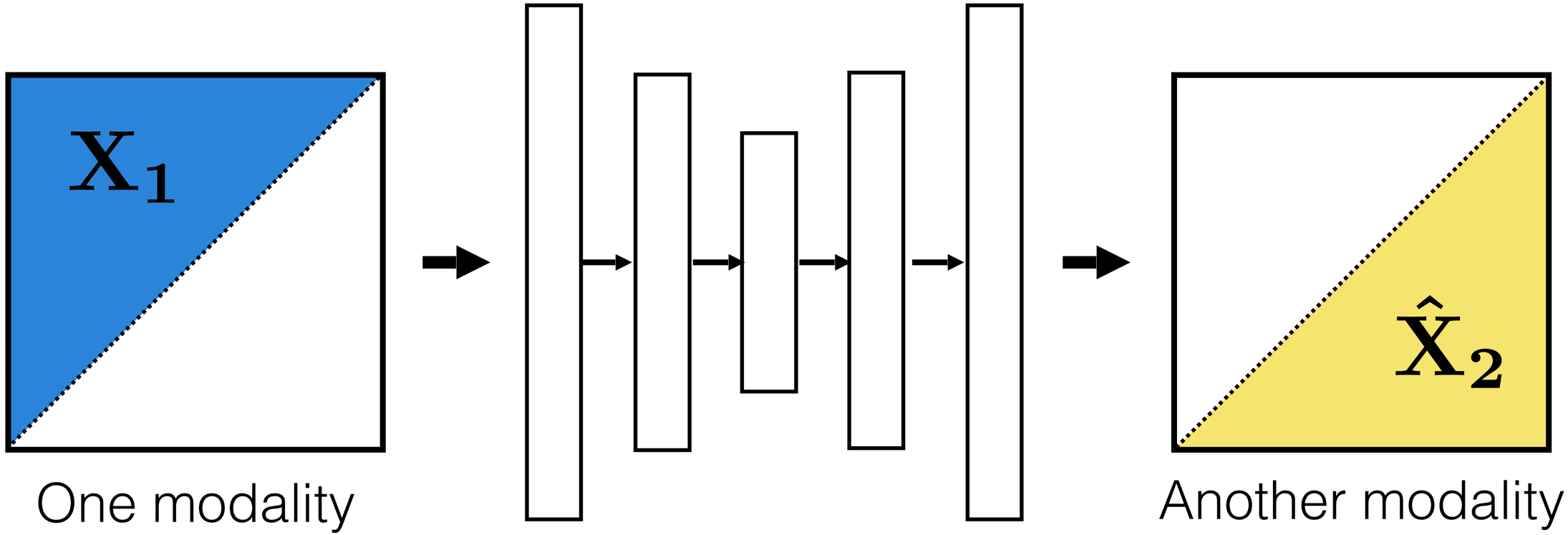
Self-Supervised

- derives label from a co-occurring input to another modality



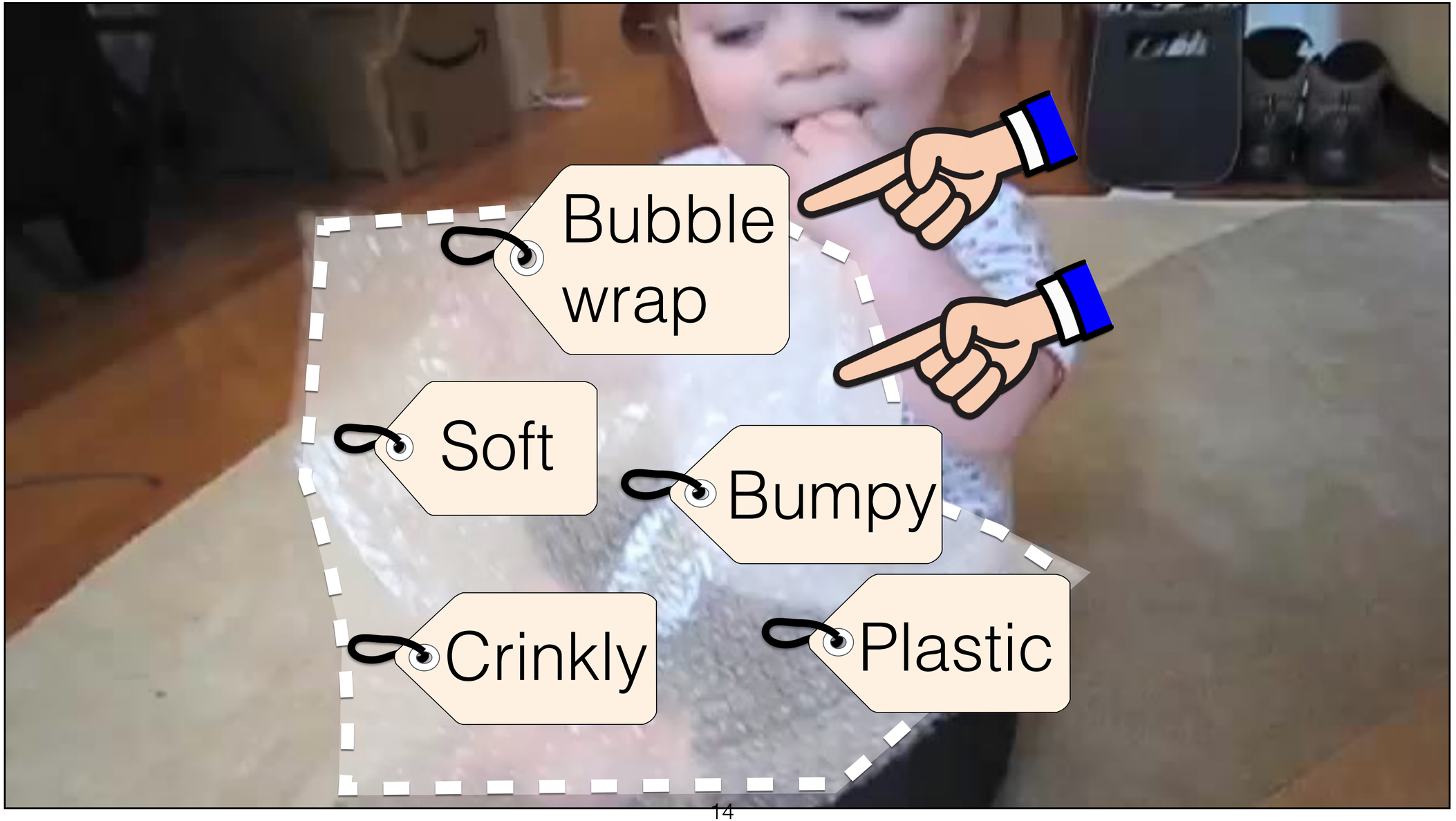
Virginia de Sa. Learning Classification with Unlabeled Data. NIPS 1994.

Multisensory self-supervision



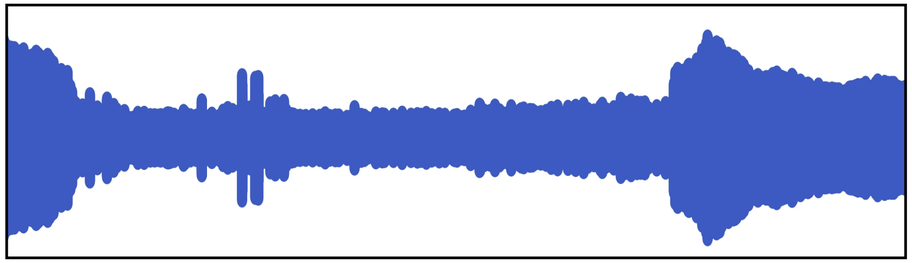
Source: Isola, Freeman, Torralba

Human supervision

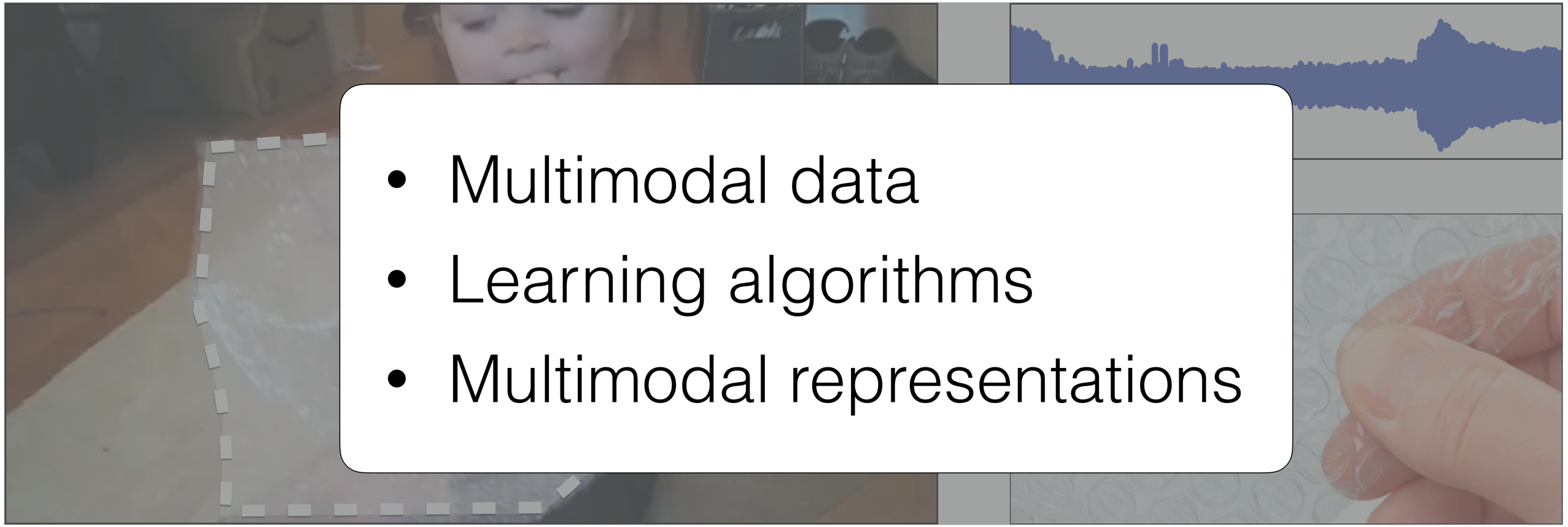


Human supervision

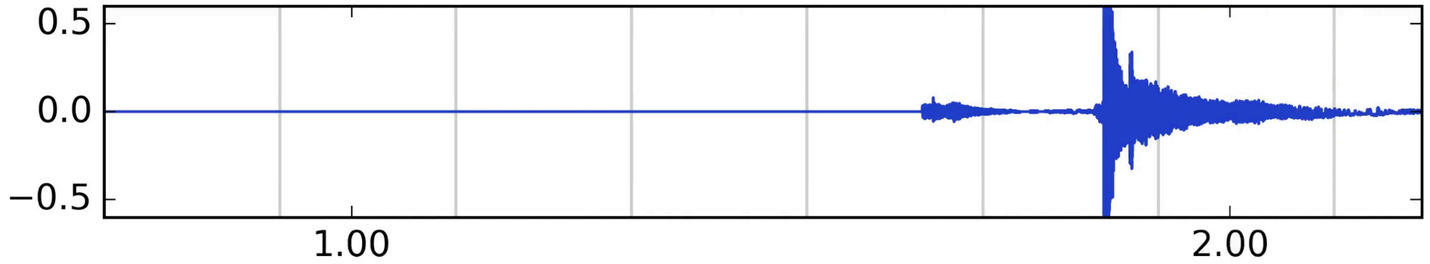
Self-supervision

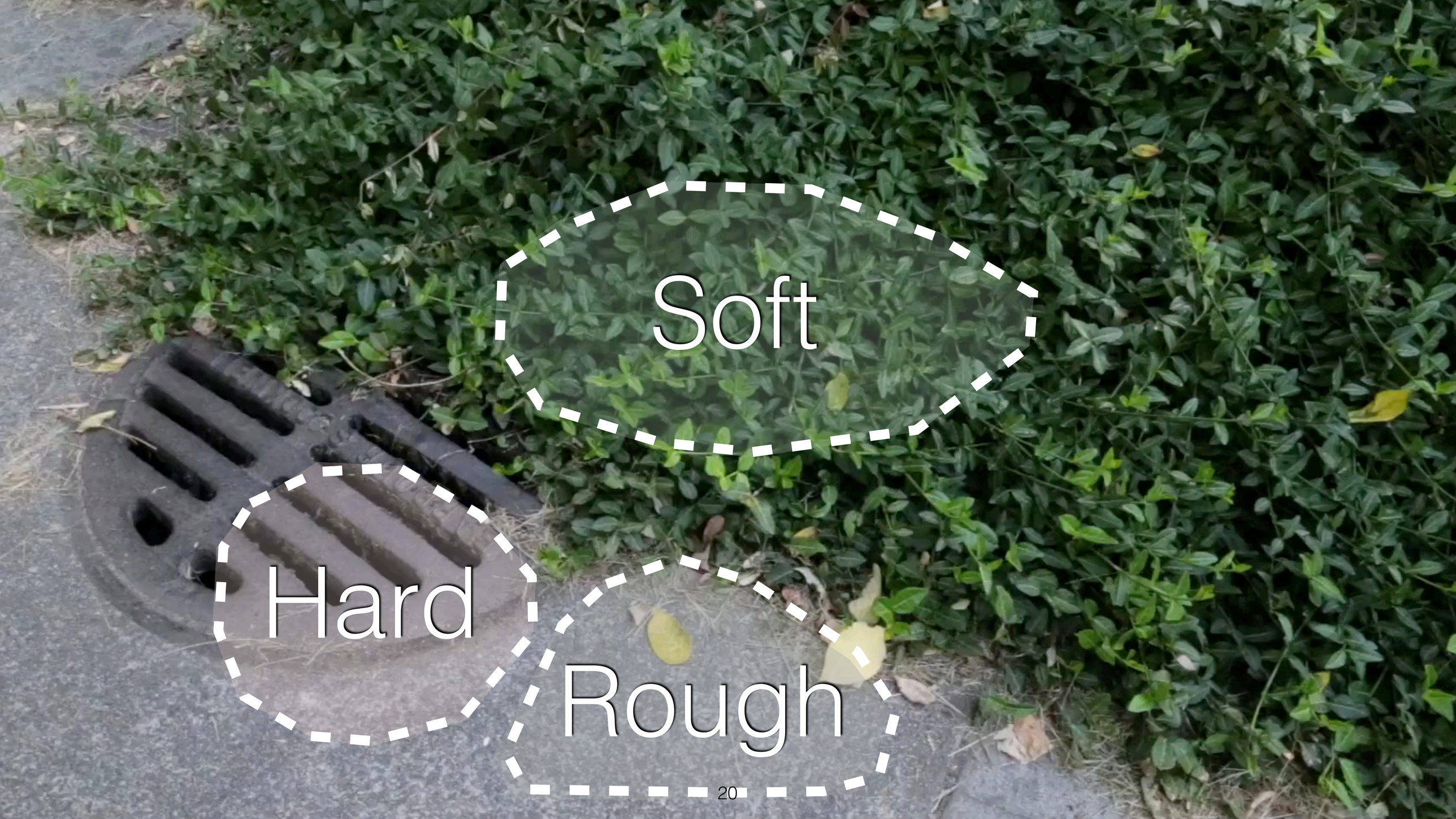


Self-supervision



- Multimodal data
- Learning algorithms
- Multimodal representations


0.5
0.0
−0.5
1.00
2.00


Soft
Hard
Rough

Learning about physical interactions

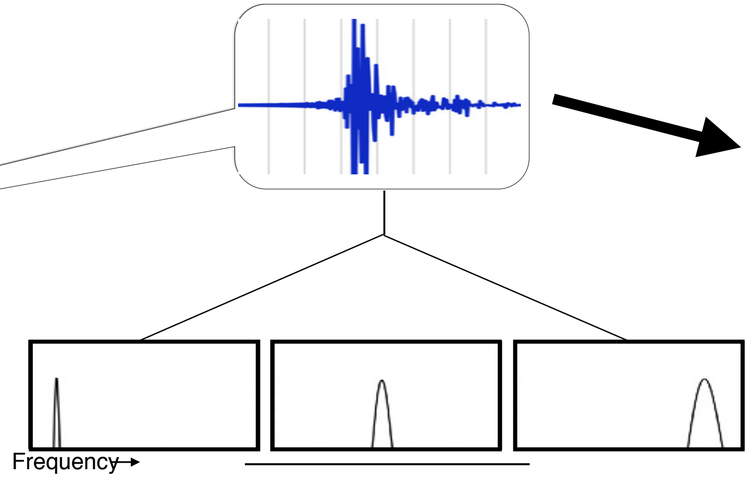
Silent video

Predicted soundtrack

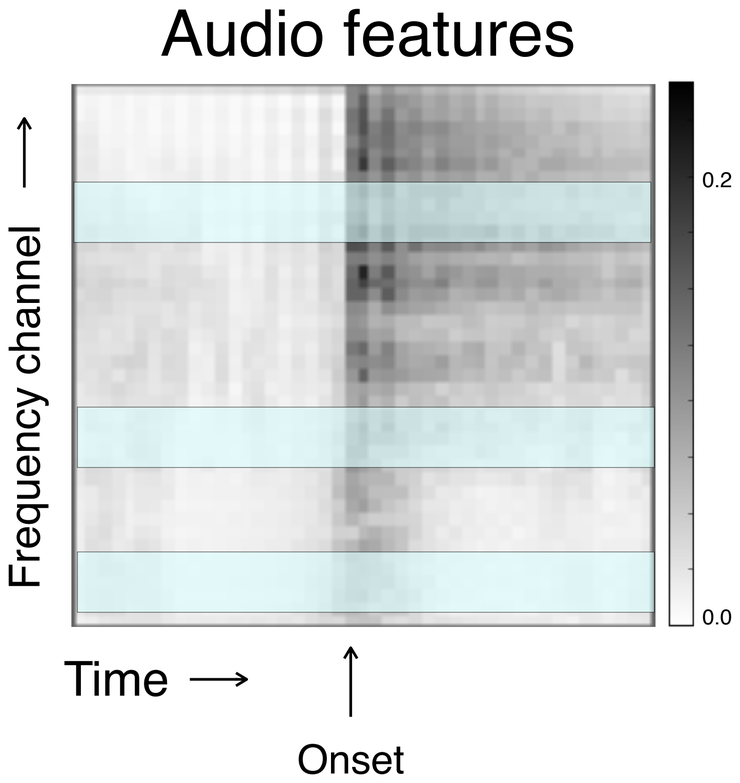
A. Owens, P. Isola, J. McDermott, A. Torralba, E. H. Adelson, W. T. Freeman. Visually Indicated Sounds. CVPR 2016.


The *Greatest Hits* Dataset: Volume 1

Can we predict physical properties from sound?



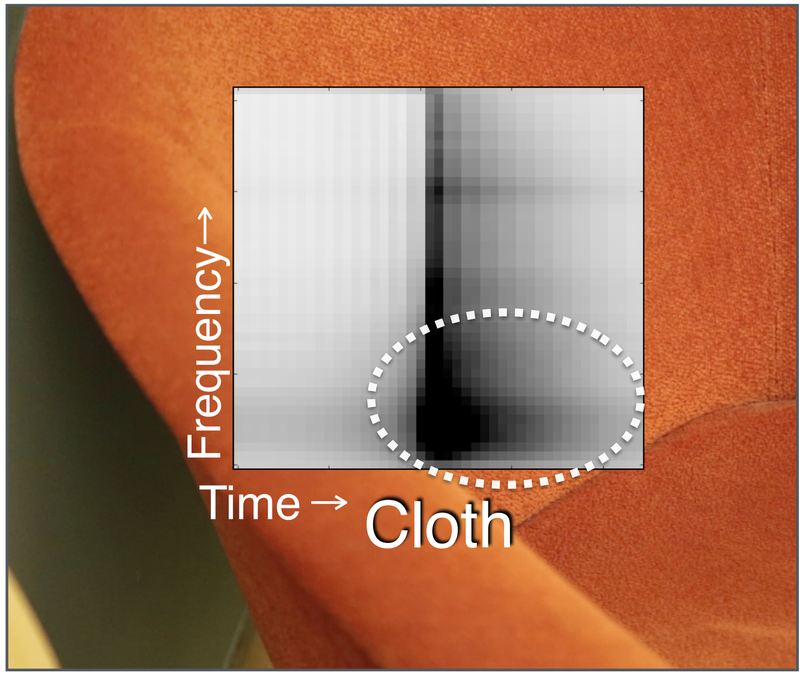
Filters tuned to different frequencies

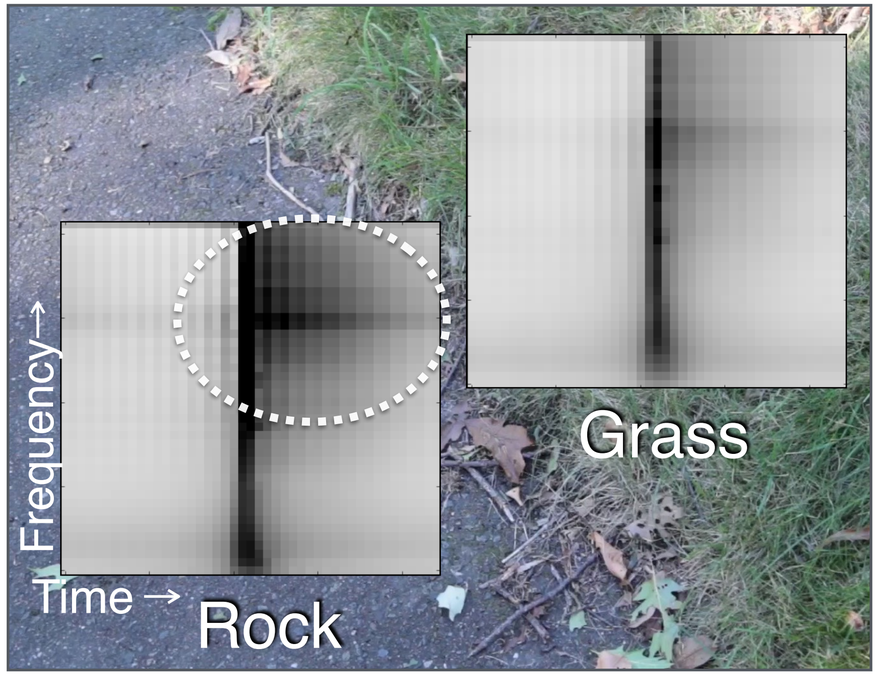


- 40 bandpass filters (+ high/low pass)

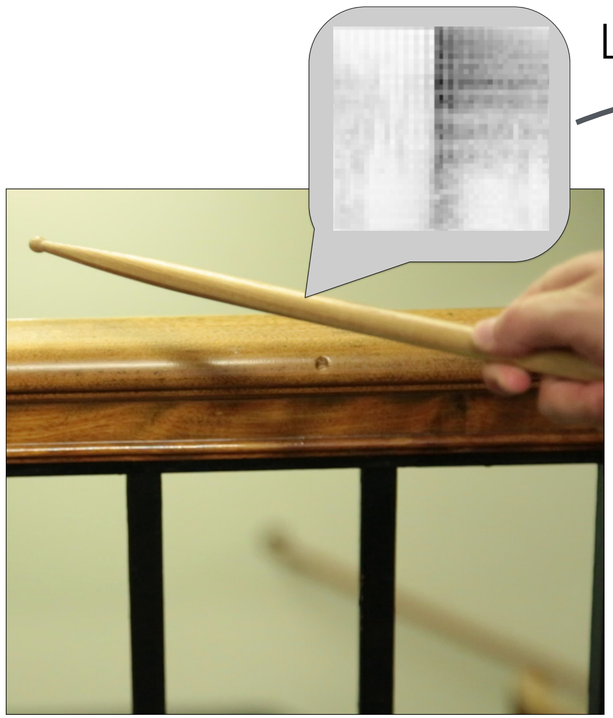
Can we predict physical properties from sound?

Mean sound features per category



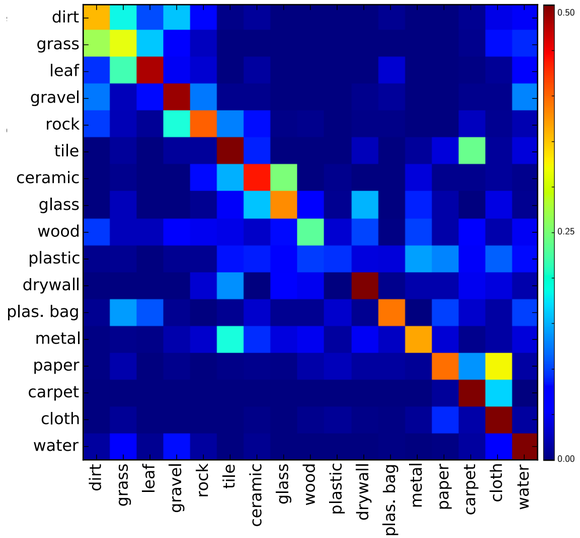


Can we predict physical properties from sound?



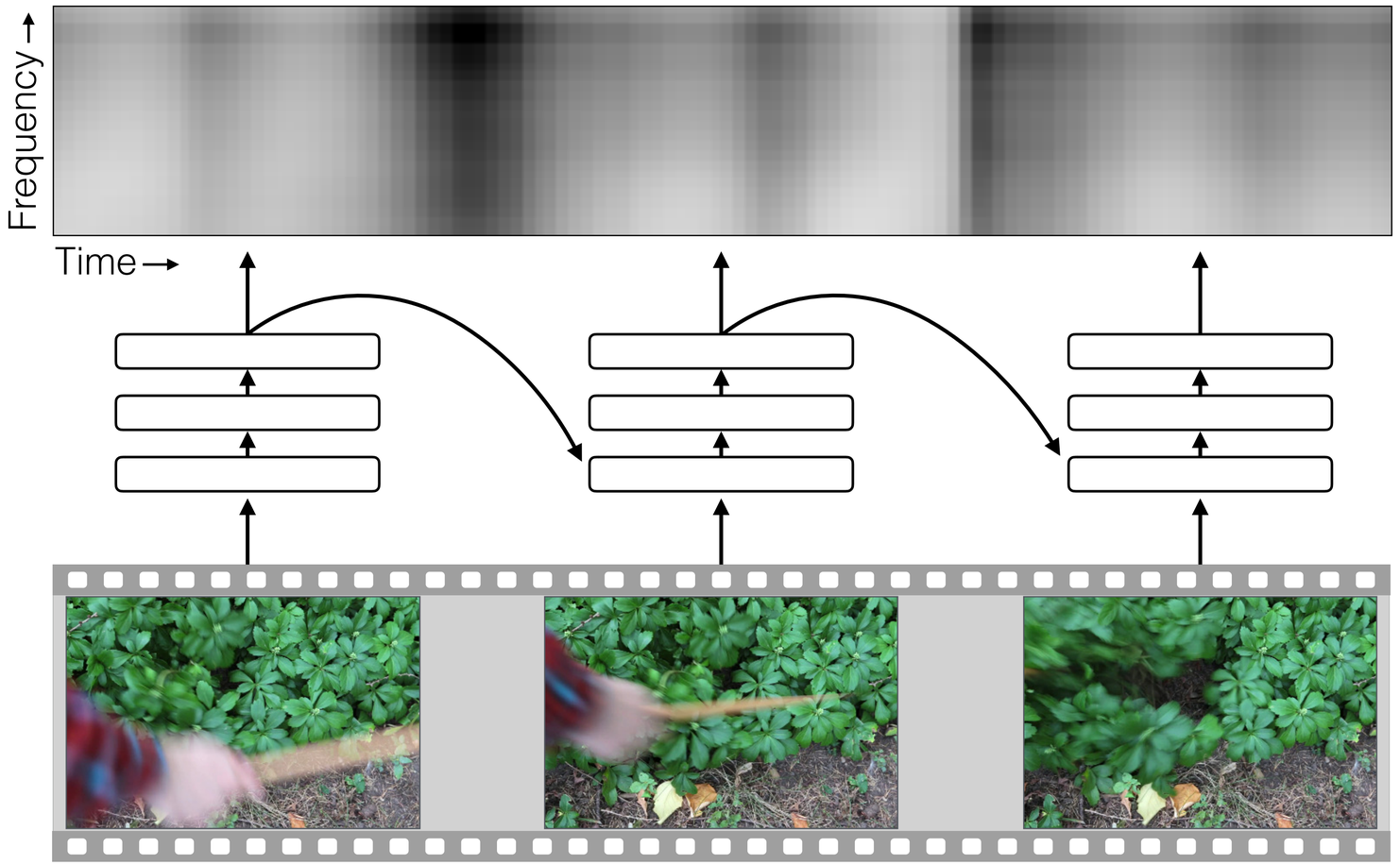
Linear classifier

“Wood”



- 46% class-averaged accuracy (chance = 6%)
- Common audio confusions:
  - {cloth, paper}, {dirt, grass}, {rock, tile}

Predicting visually indicated sounds

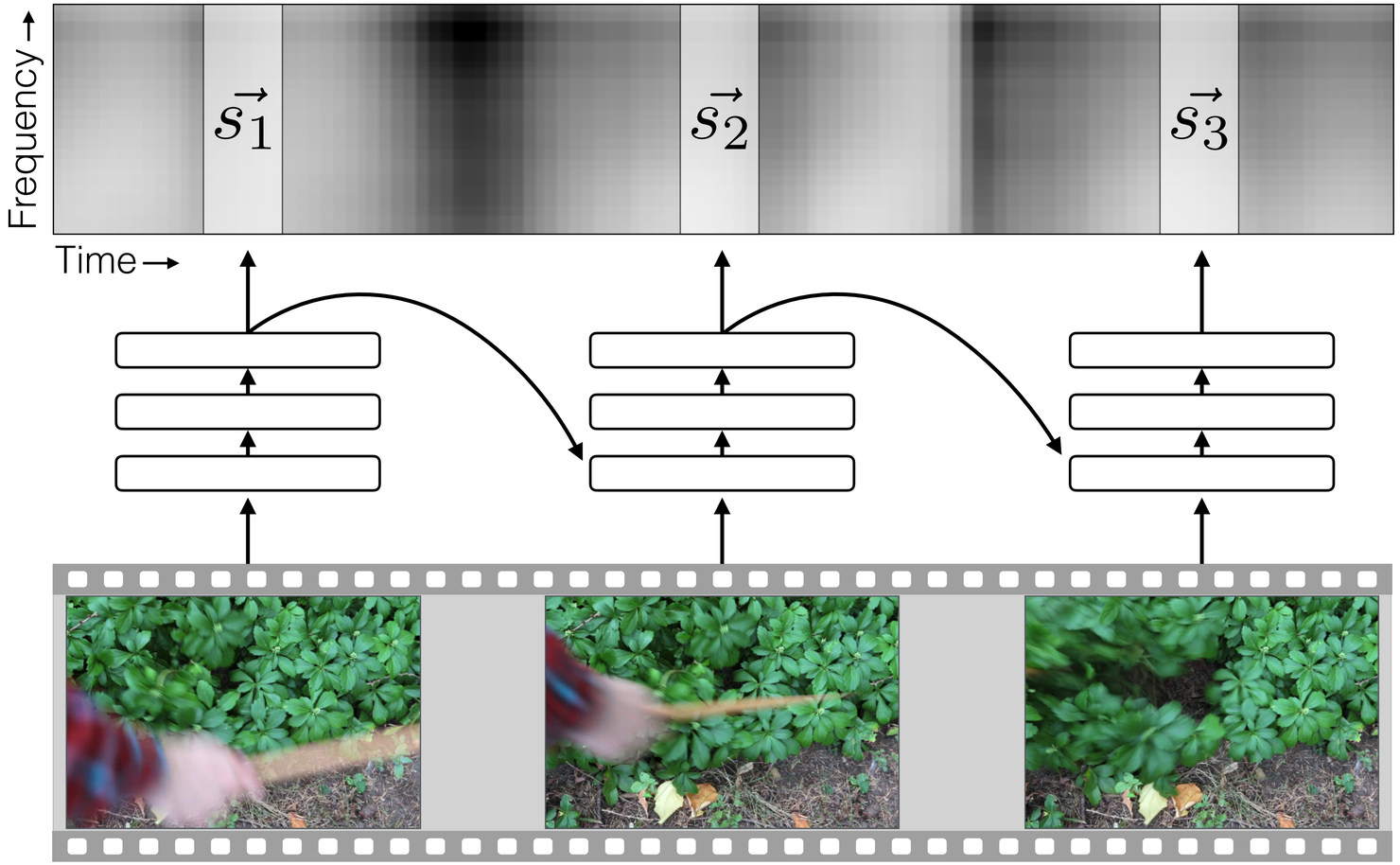


Audio features

Neural network

Images

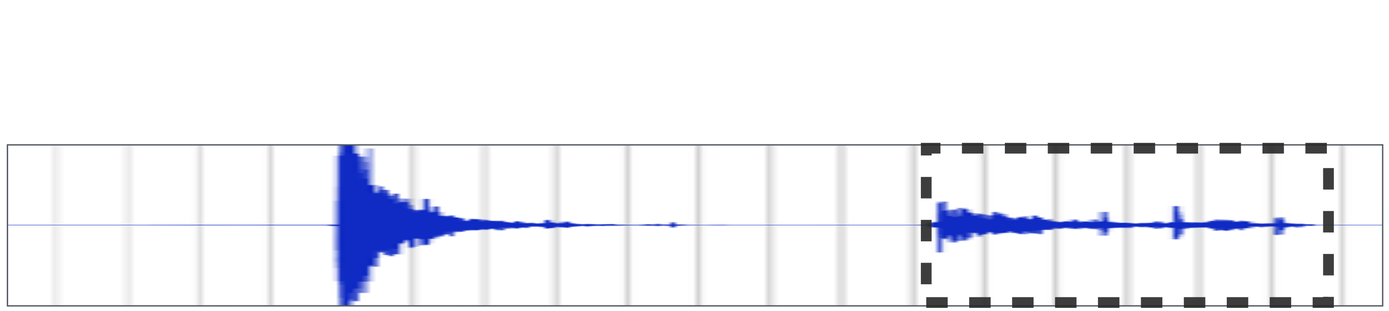
Predicting visually indicated sounds

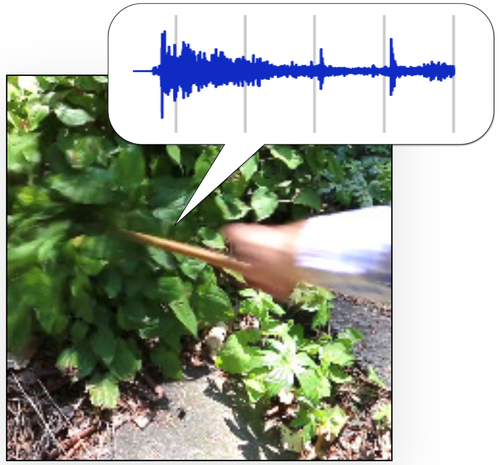


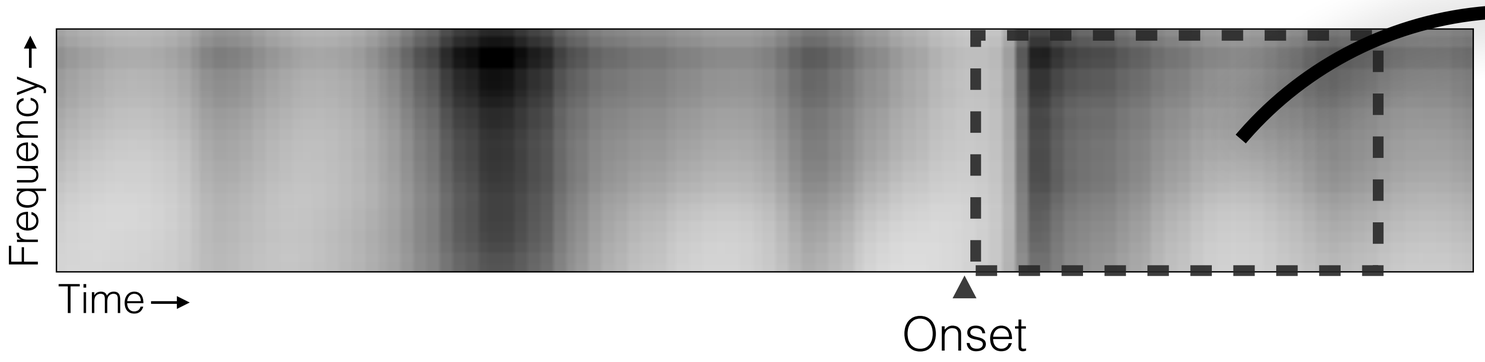
Ground truth

$$\sum_{t=1}^{T} \|\vec{s}_t - \tilde{\vec{s}}_t\|$$

Predicting visually indicated sounds

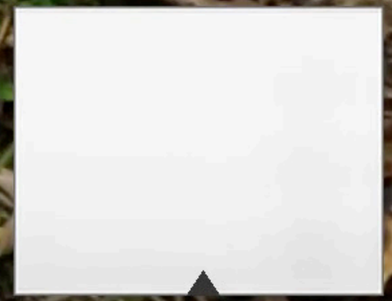






Sound database

Predicted sound

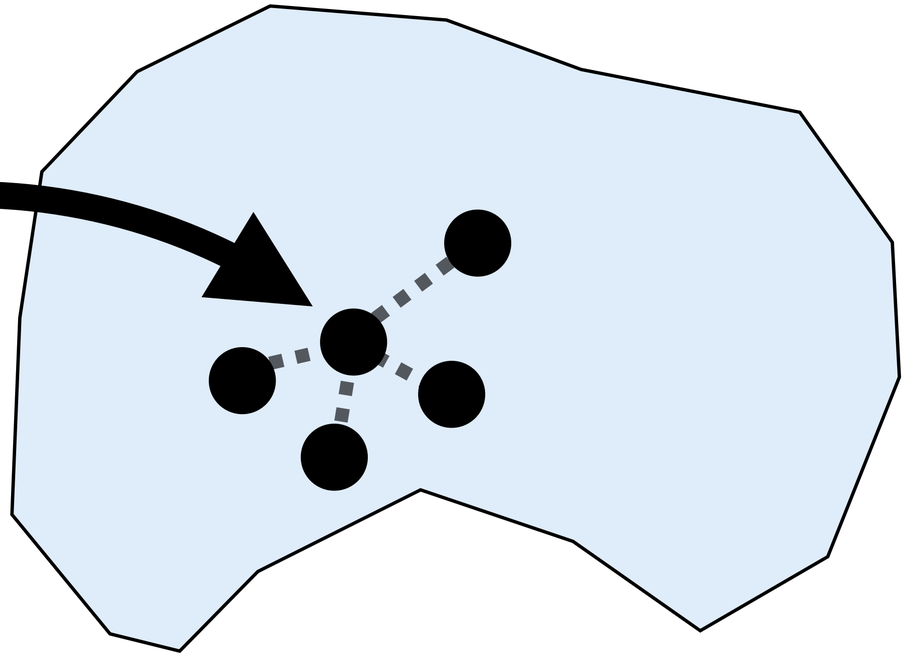


Predicted sound

Predicted sound

Failure case

Translating between modalities



Audio space

Translating between modalities

Audio nearest neighbors have similar material properties

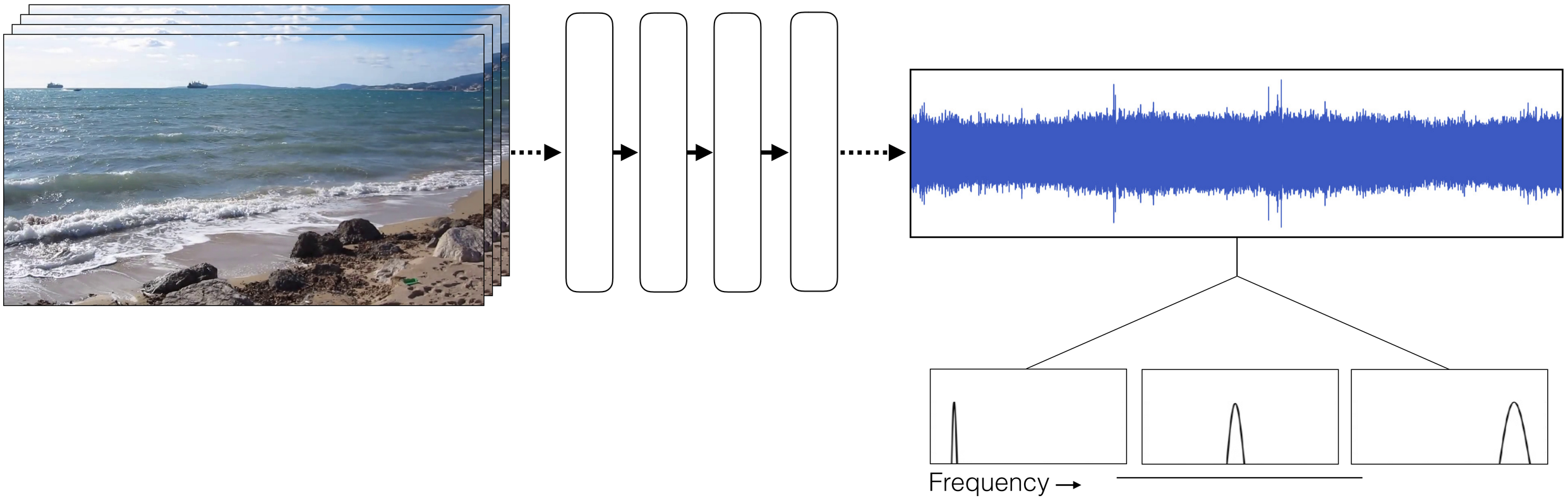
Active vs. passive perception

A. Owens, J. Wu, J. H. McDermott, A. Torralba, W. T. Freeman.
Ambient Sound Provides Supervision for Visual Learning. ECCV 2016.

Active vs. passive perception

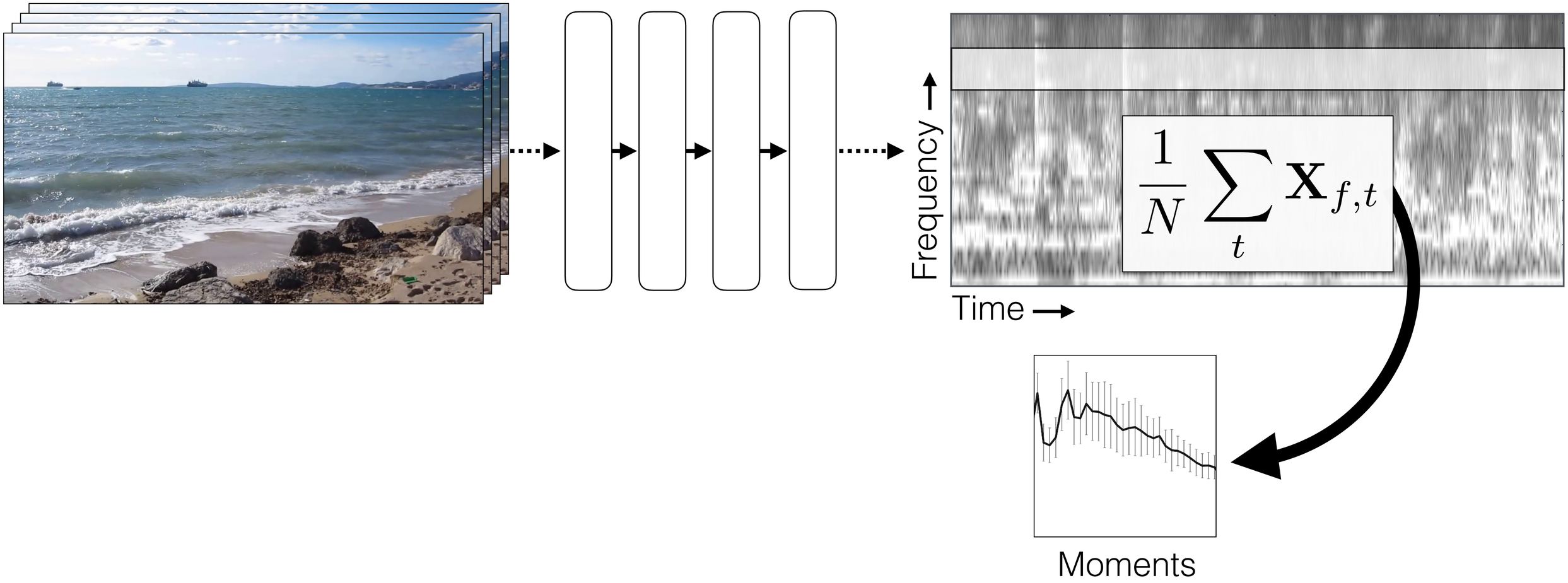
A. Owens, J. Wu, J. H. McDermott, A. Torralba, W. T. Freeman.
Ambient Sound Provides Supervision for Visual Learning. ECCV 2016.

Predicting ambient sound

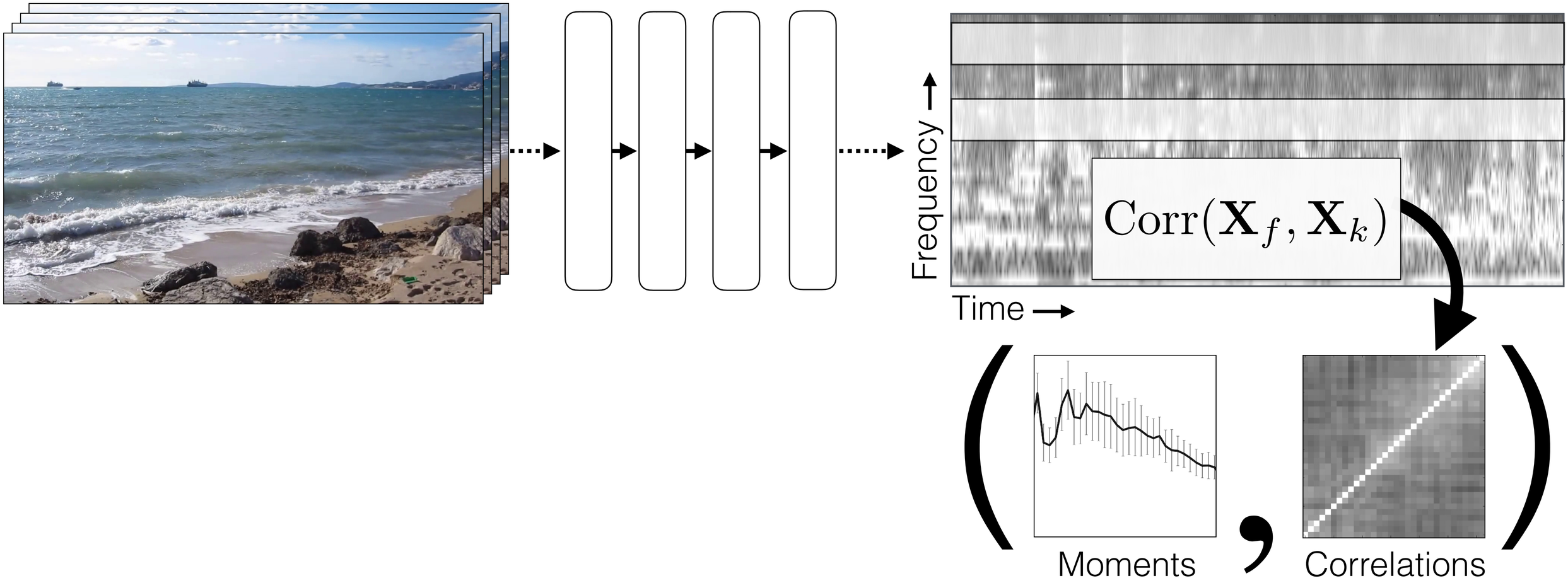


Filters tuned to different frequencies

Predicting ambient sound

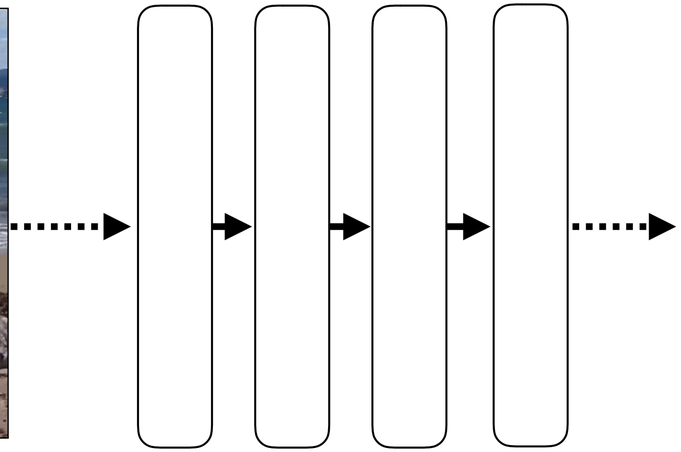


Predicting ambient sound

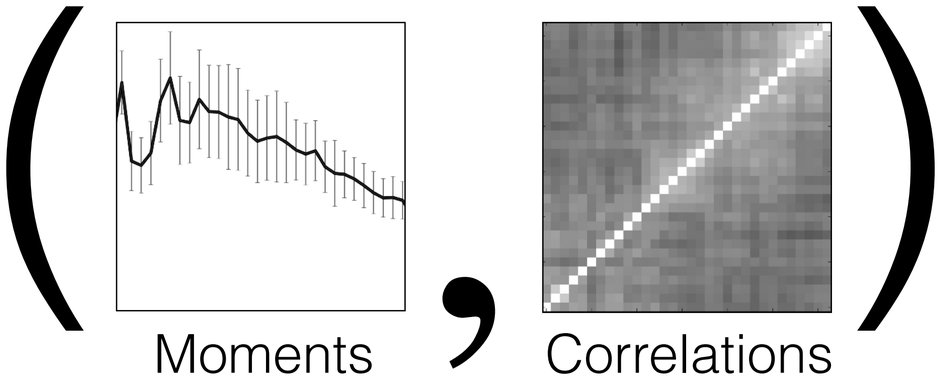


Predicting ambient sound

Video frame



Convolutional network

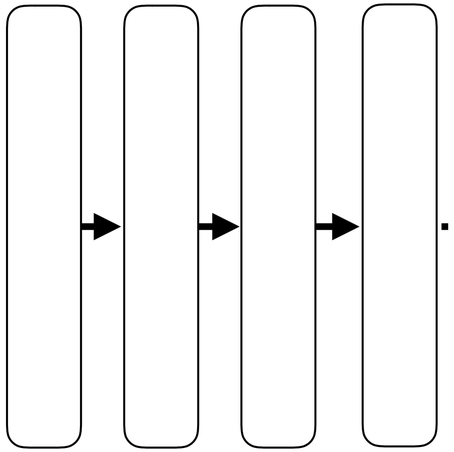


Sound texture

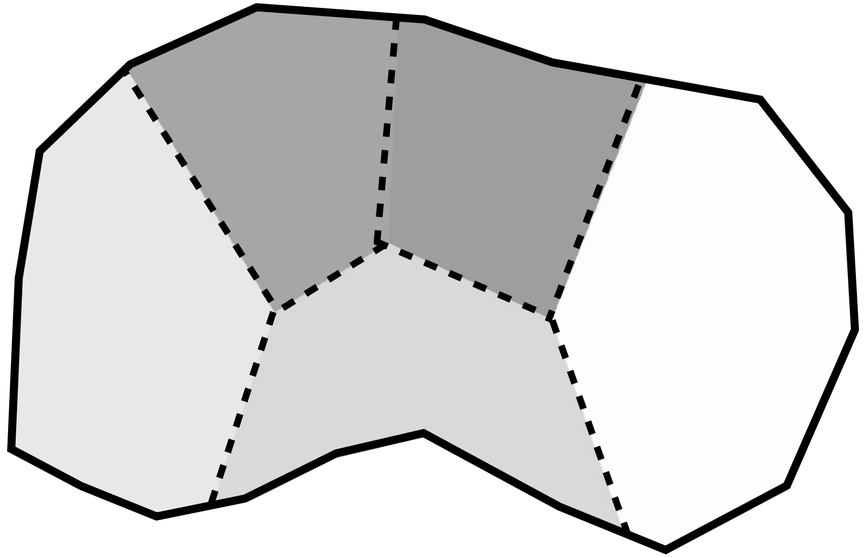
(McDermott & Simoncelli 2011)

Predicting ambient sound

Video frame



Convolutional network



$p(S \mid I)$

K-means partitioning

Audio is invariant to visual transformations

Image space

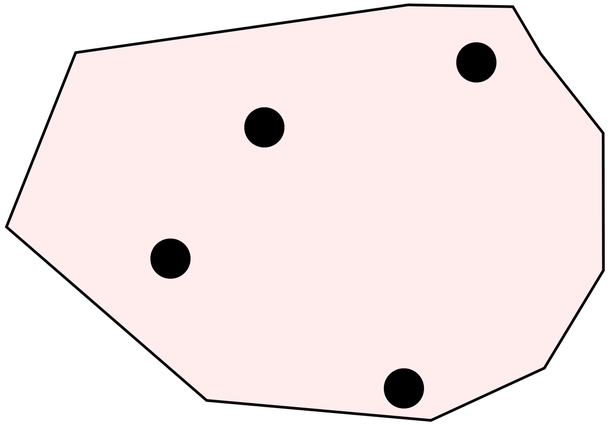
Audio space

Visual model

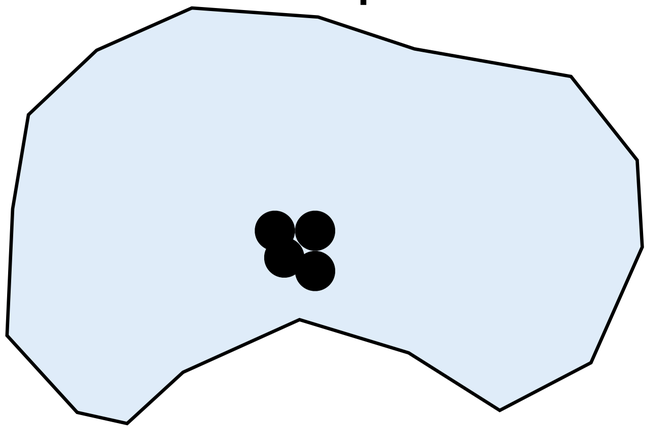
Common causes

Audio is invariant to visual transformations

Image space

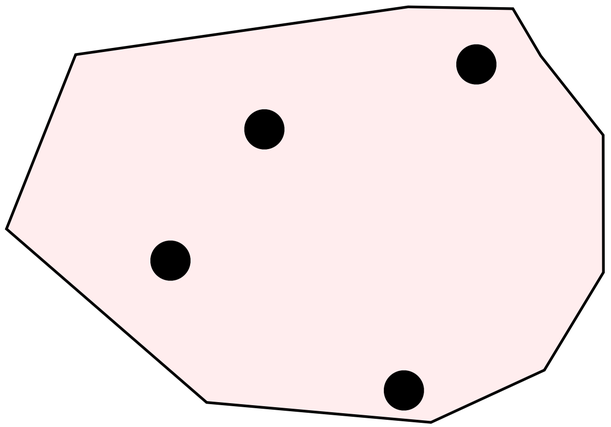


Audio space

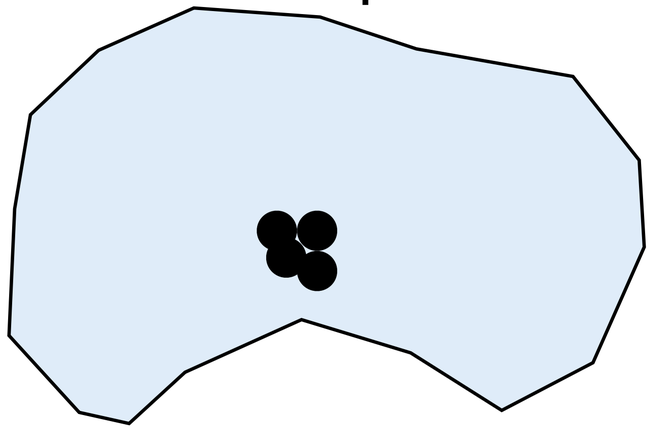


Audio is invariant to visual transformations

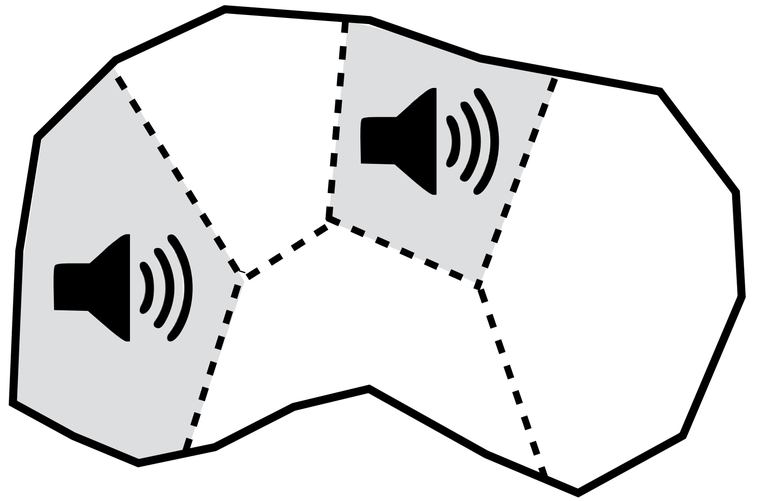
Image space



Audio space



What did the model learn?

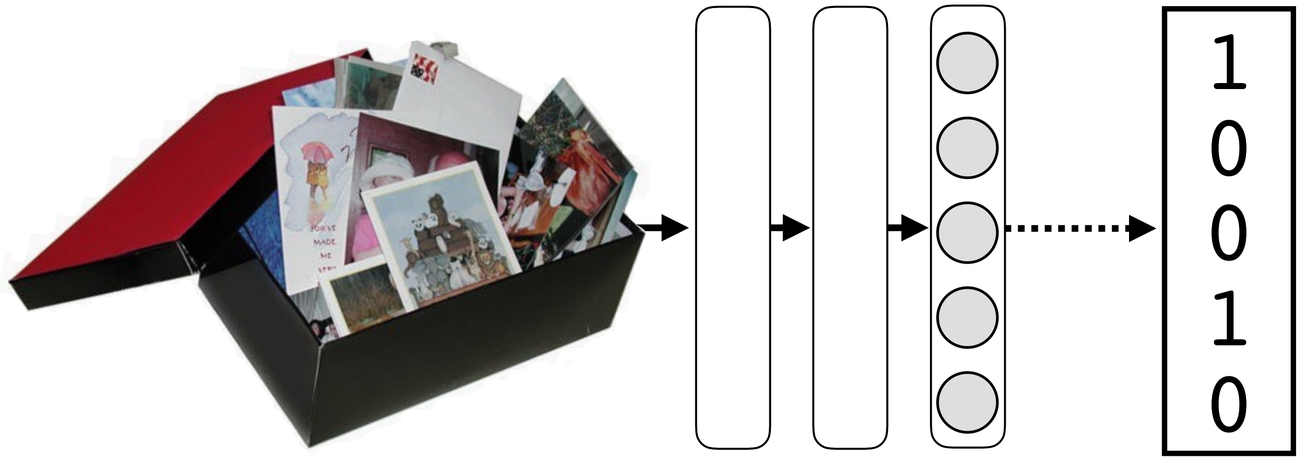


Audio space

$$p(S \mid \quad)$$

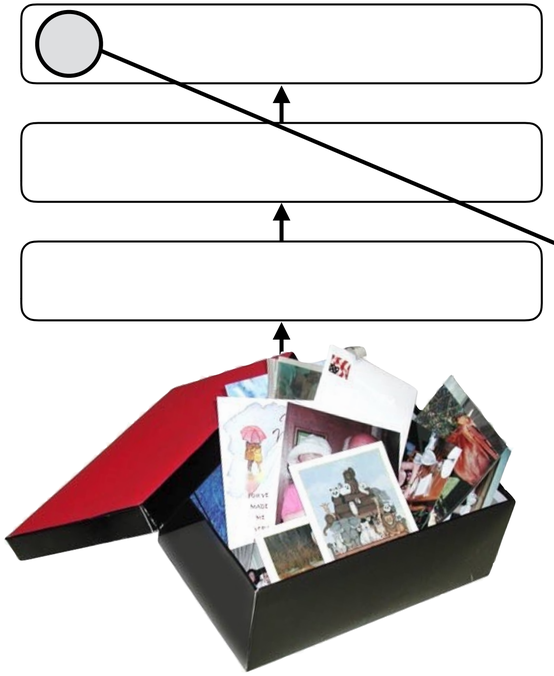
Class activation map (Zhou 2016)

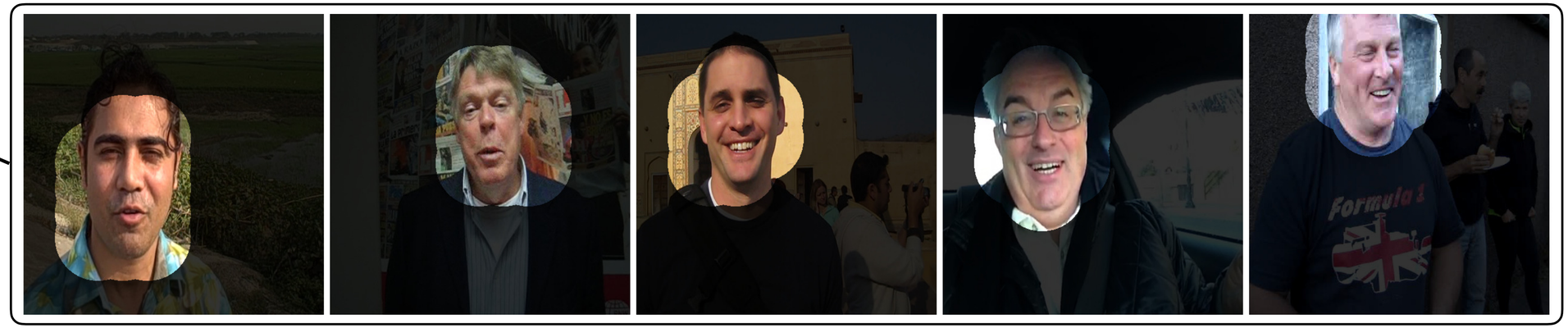
What did the model learn?



What did the model learn?

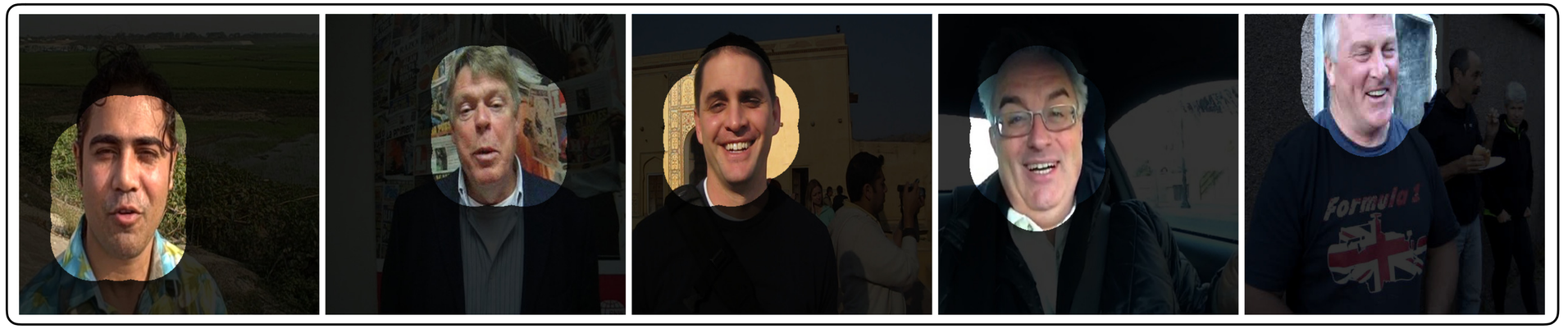
Unit #90 of 256

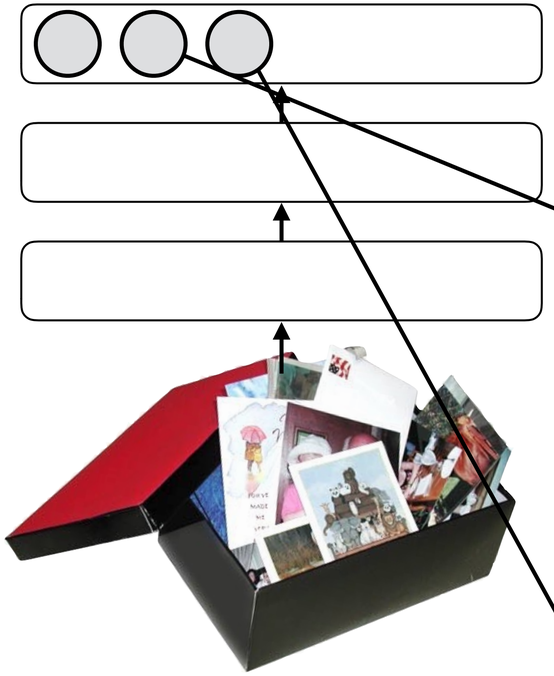


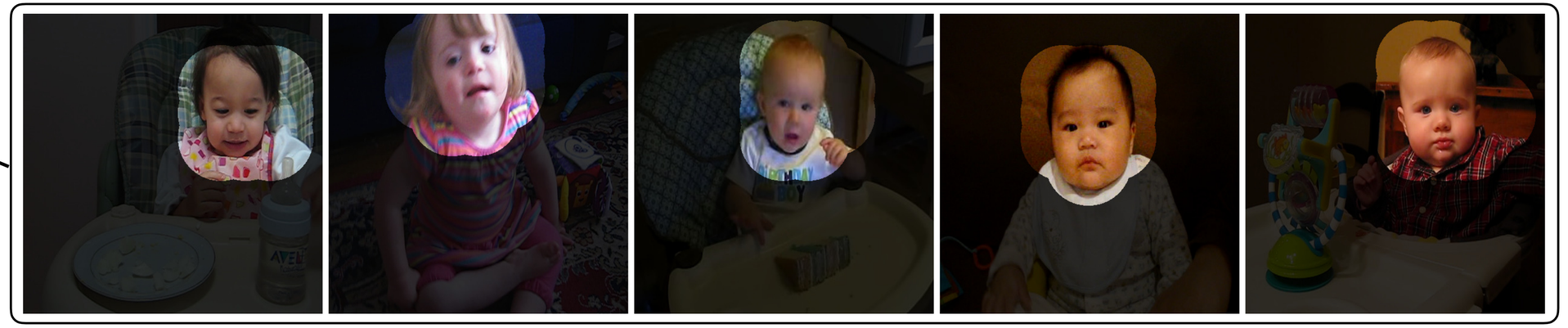


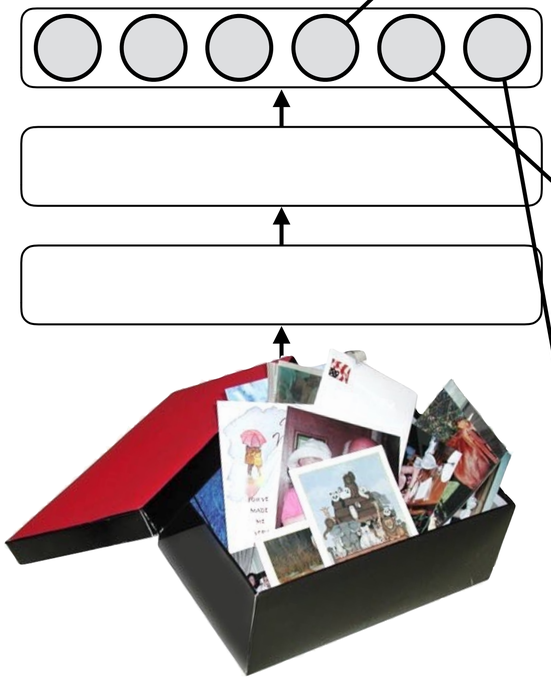
Strongest responses in dataset

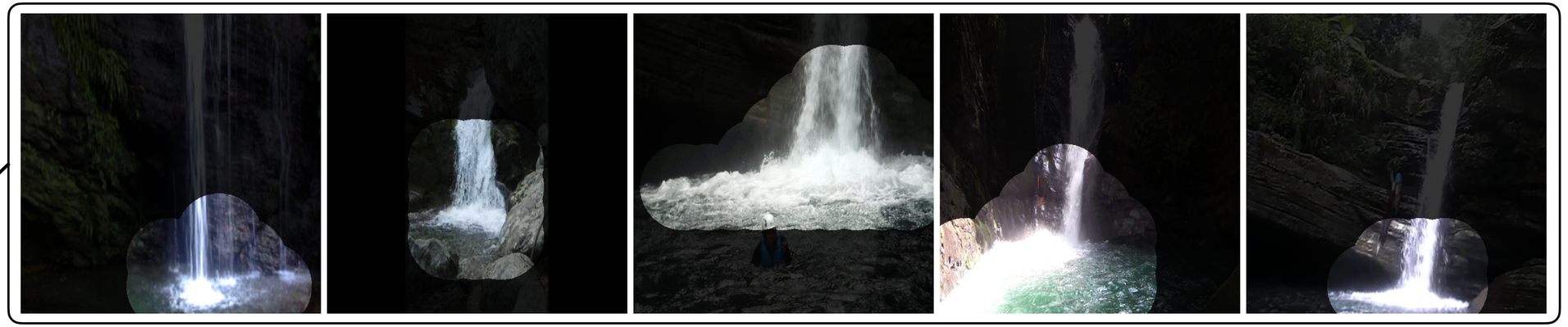
Visualization method from (Zhou 2015)



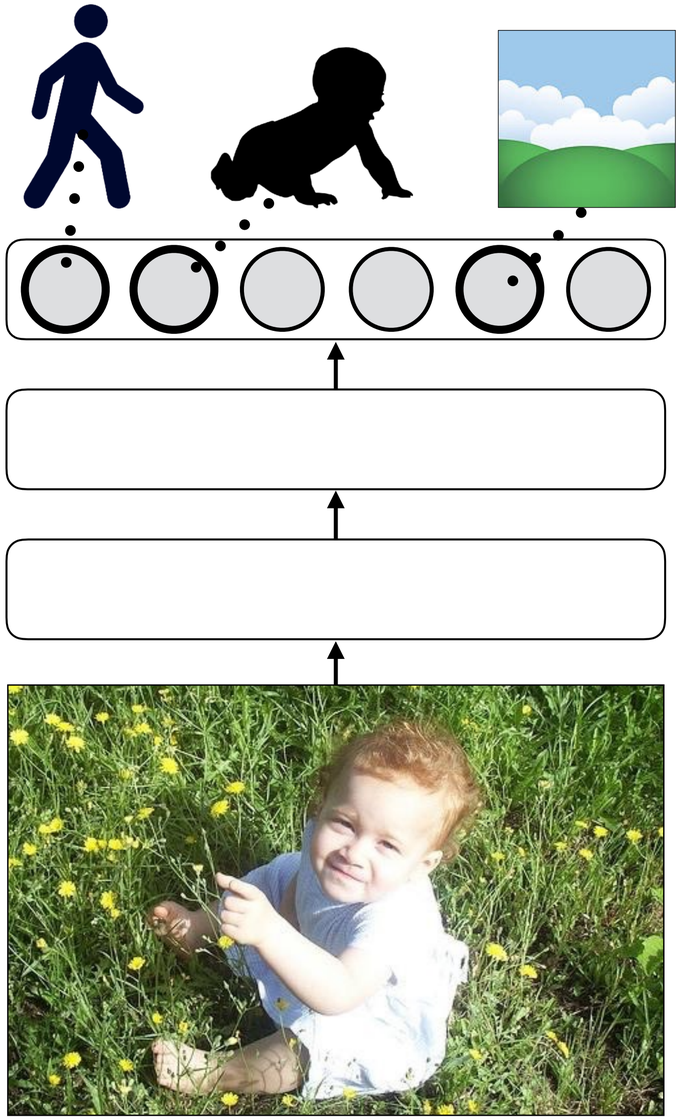




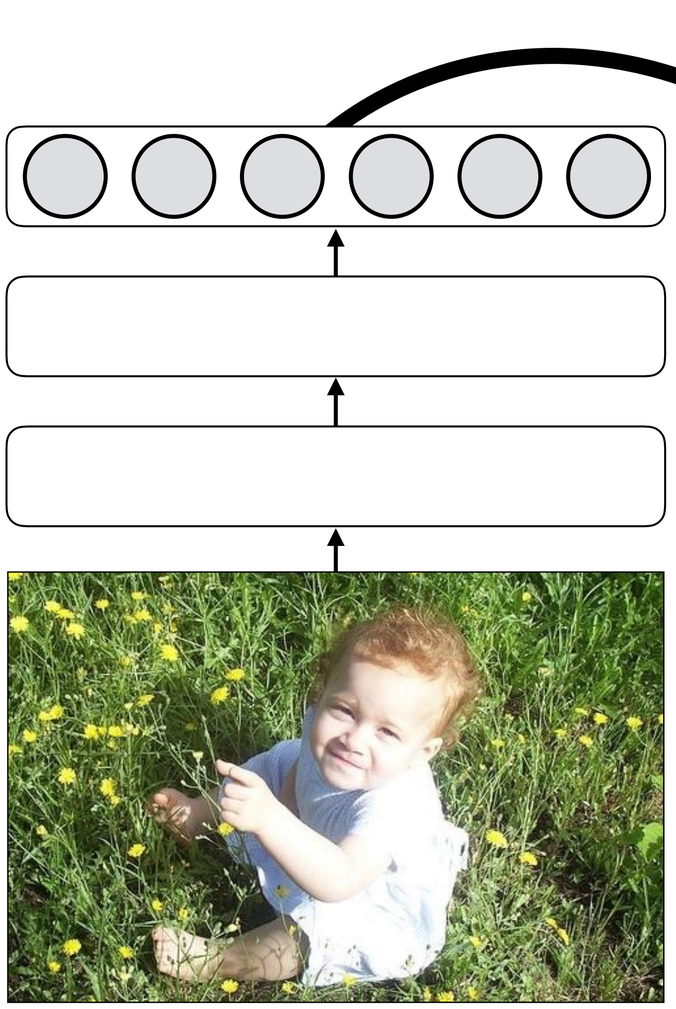




What did the model learn?



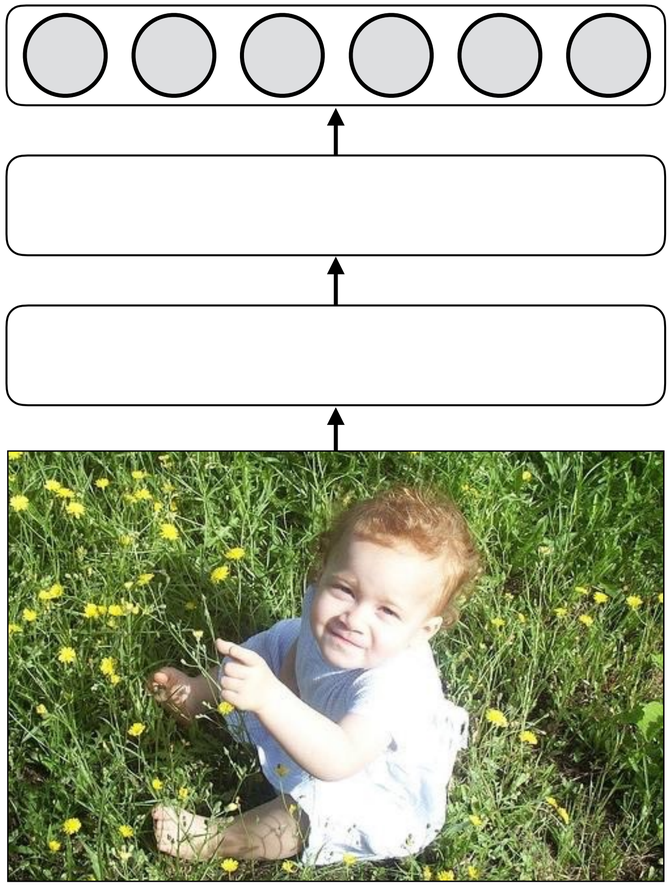
What did the model learn?

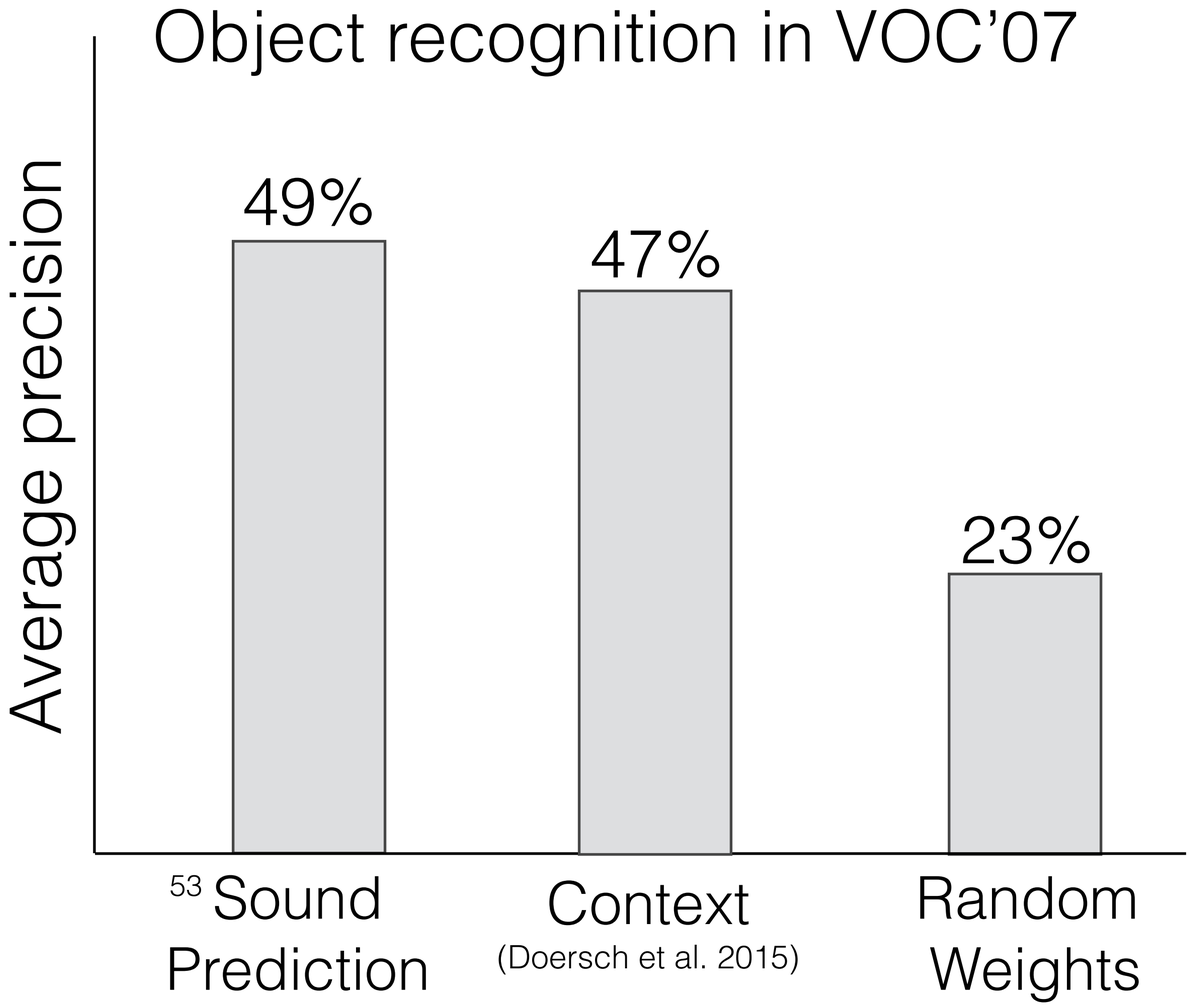


Linear Classifier

“Person”

What did the model learn?

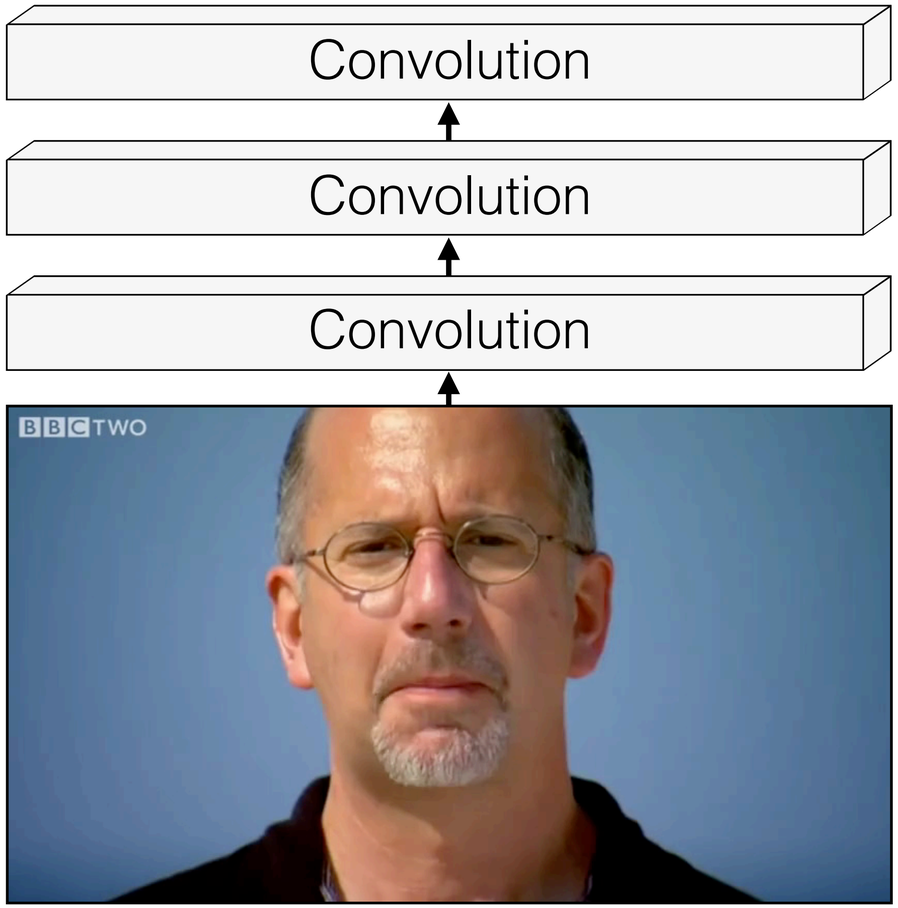




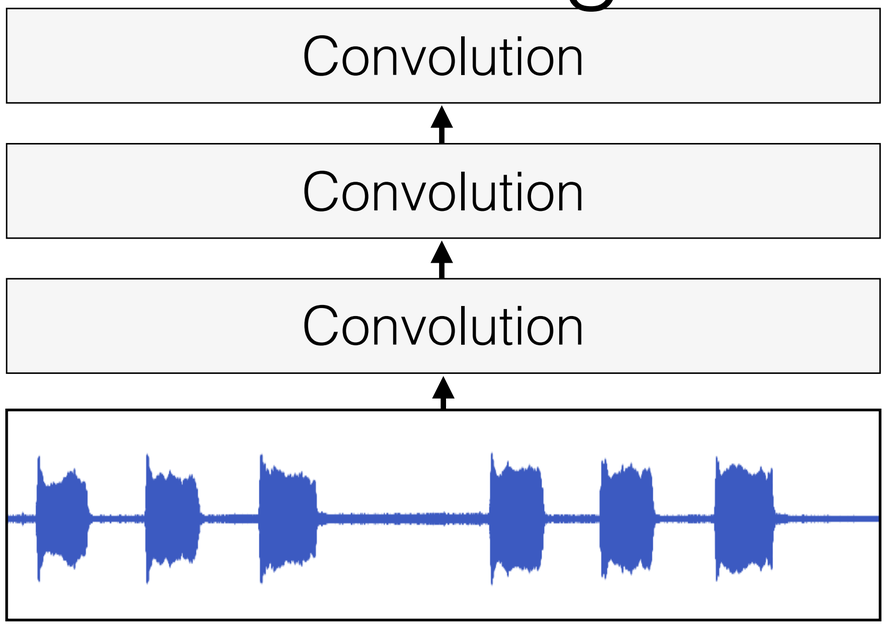
Learning a multimodal representation

Single-modality representations

Vision



Hearing

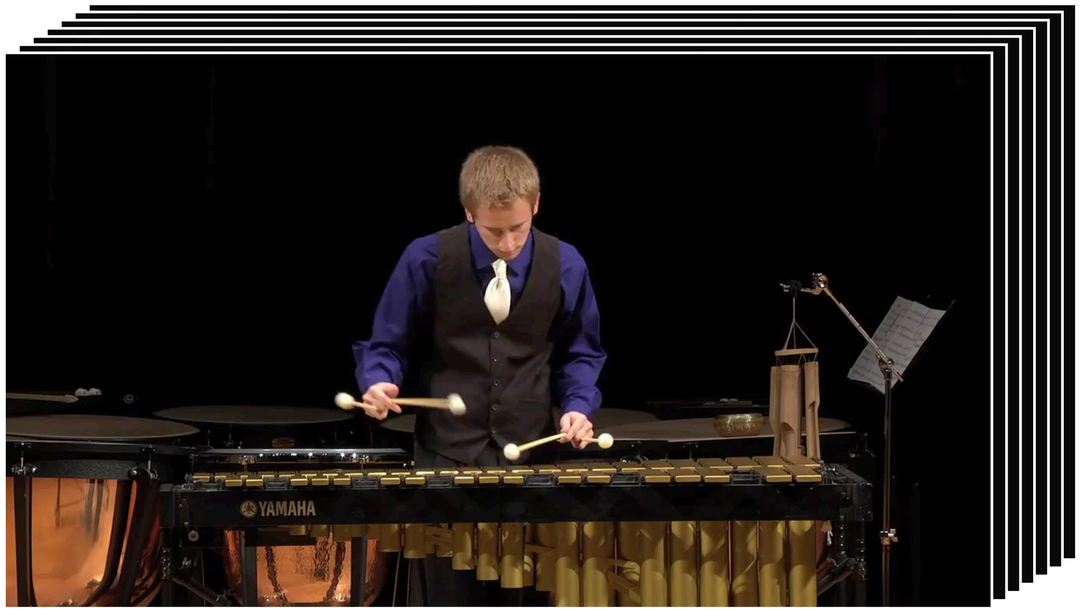


A. Owens, A. A. Efros. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features. ECCV 2018.

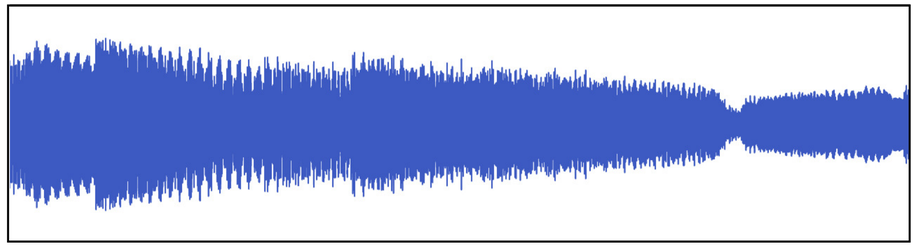
Idea: train a model to find audio-visual correspondences in video.


How do we get ground-truth correspondences?

Learning audio-visual correspondences



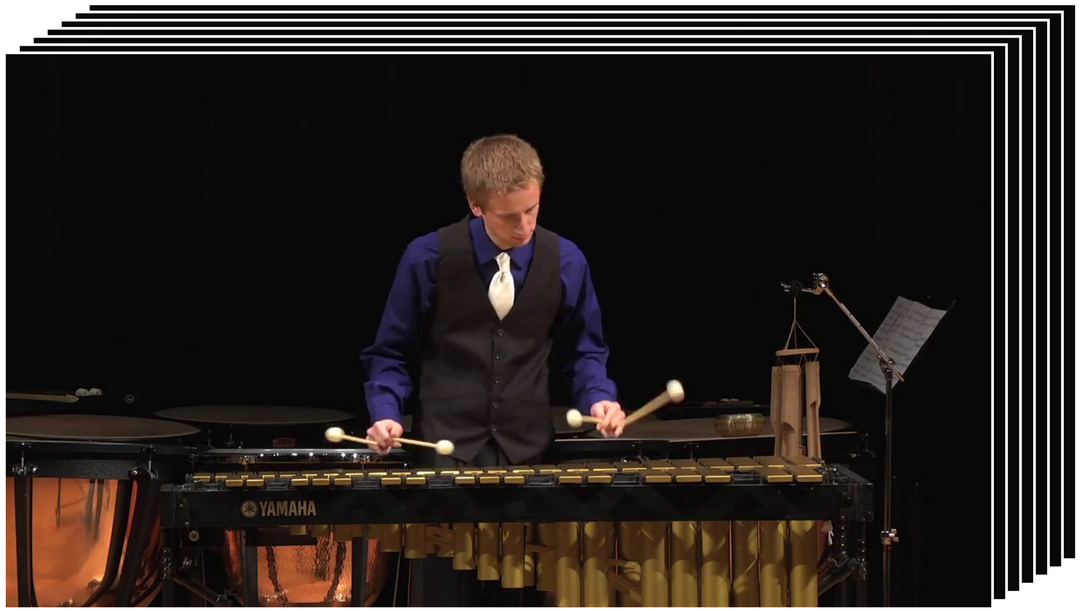
,



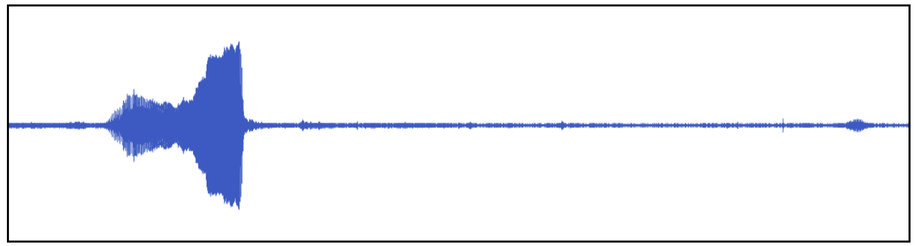
→ **real** or fake ?

Related work: L^3-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

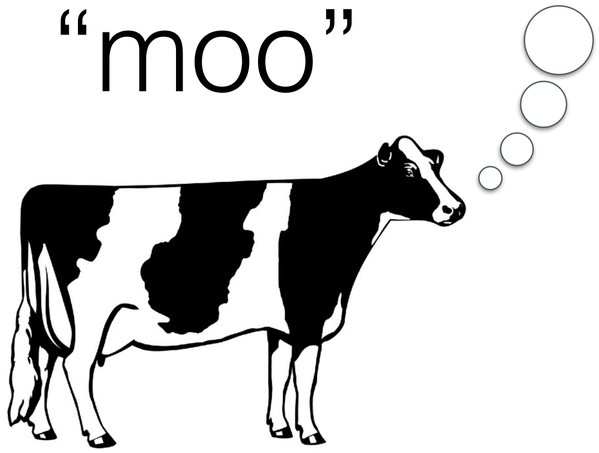
Learning audio-visual correspondences



,



“moo”



→ real or **fake**?

Related work: L^3-net (Arandjelović & Zisserman 2017), AVTS (Korbar et al. 2018)
Noise-contrastive estimation (Gutmann & Hyvarinen 2010)

Idea #1: random pairs

,

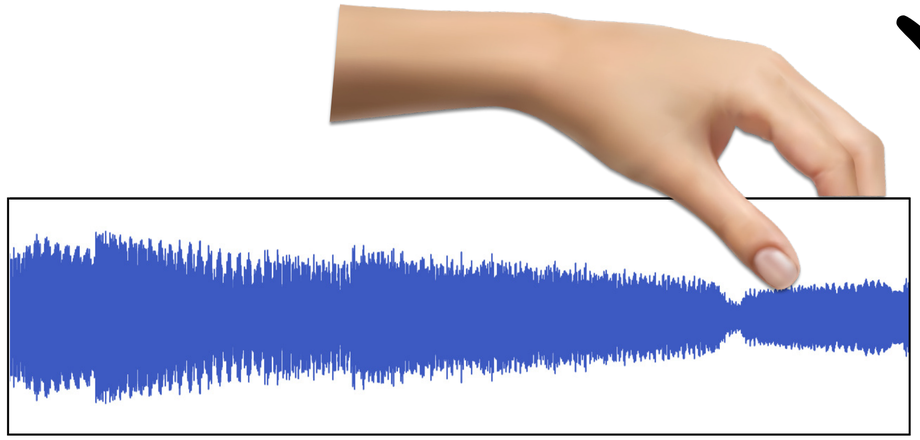
(Arandjelović & Zisserman 2017)

Idea #1: random pairs

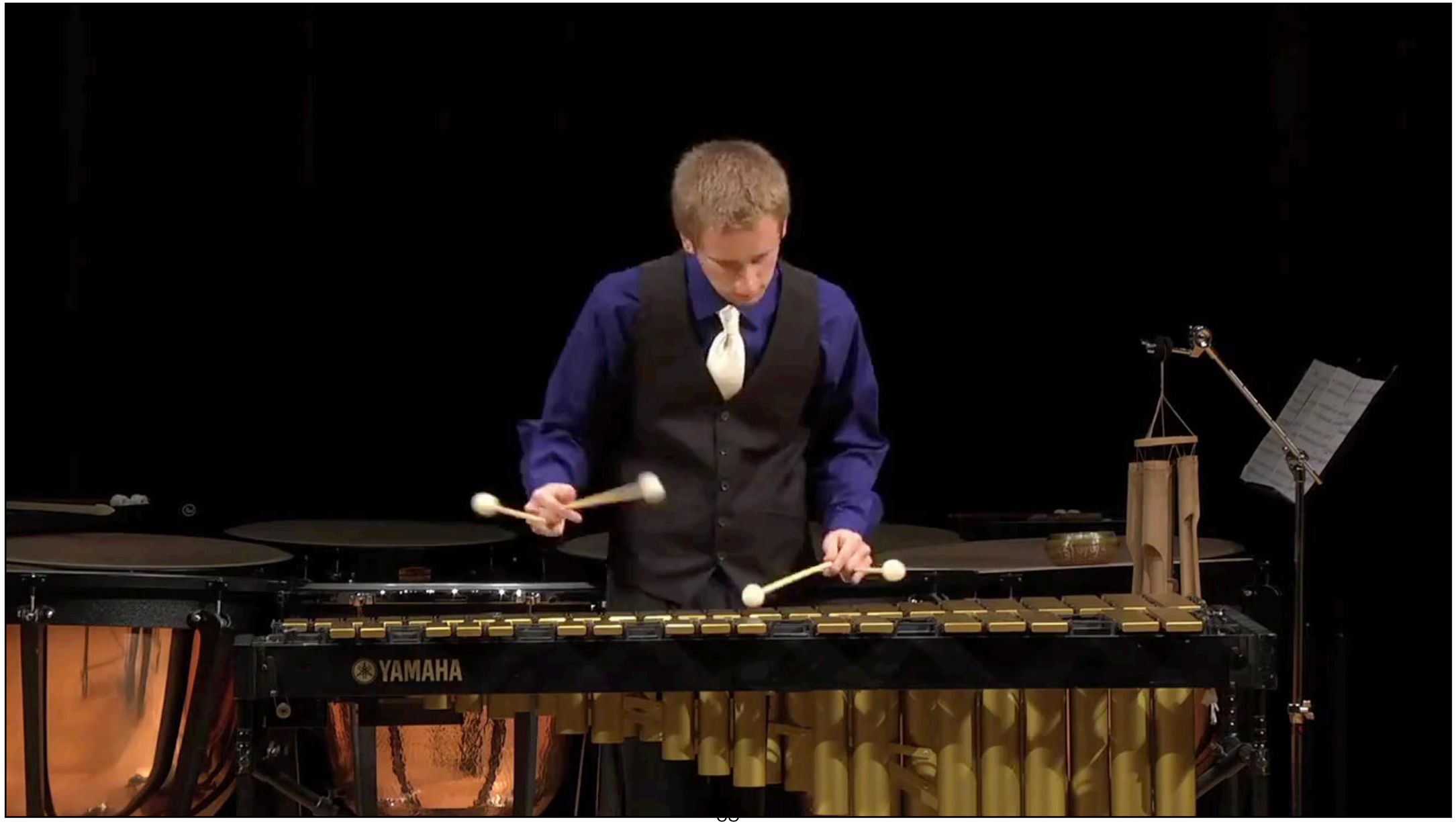
Too easy! Doesn’t require motion analysis.

Idea #2: time-shifted pairs

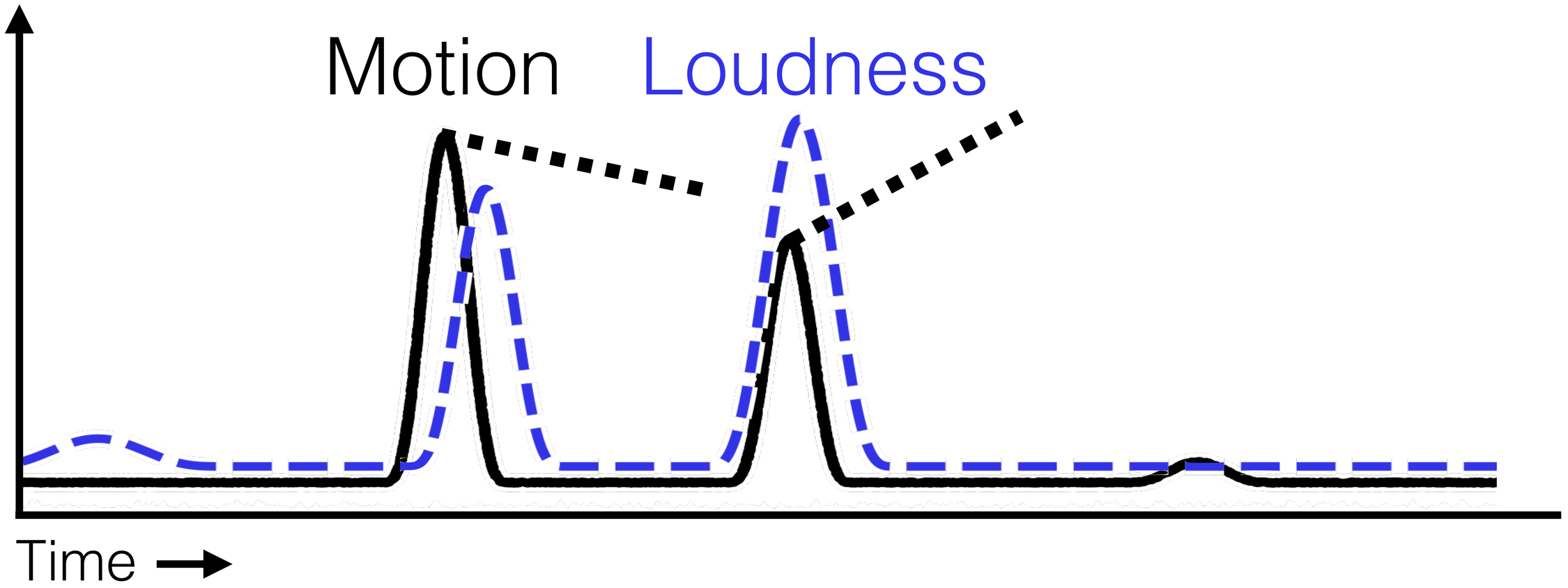
,



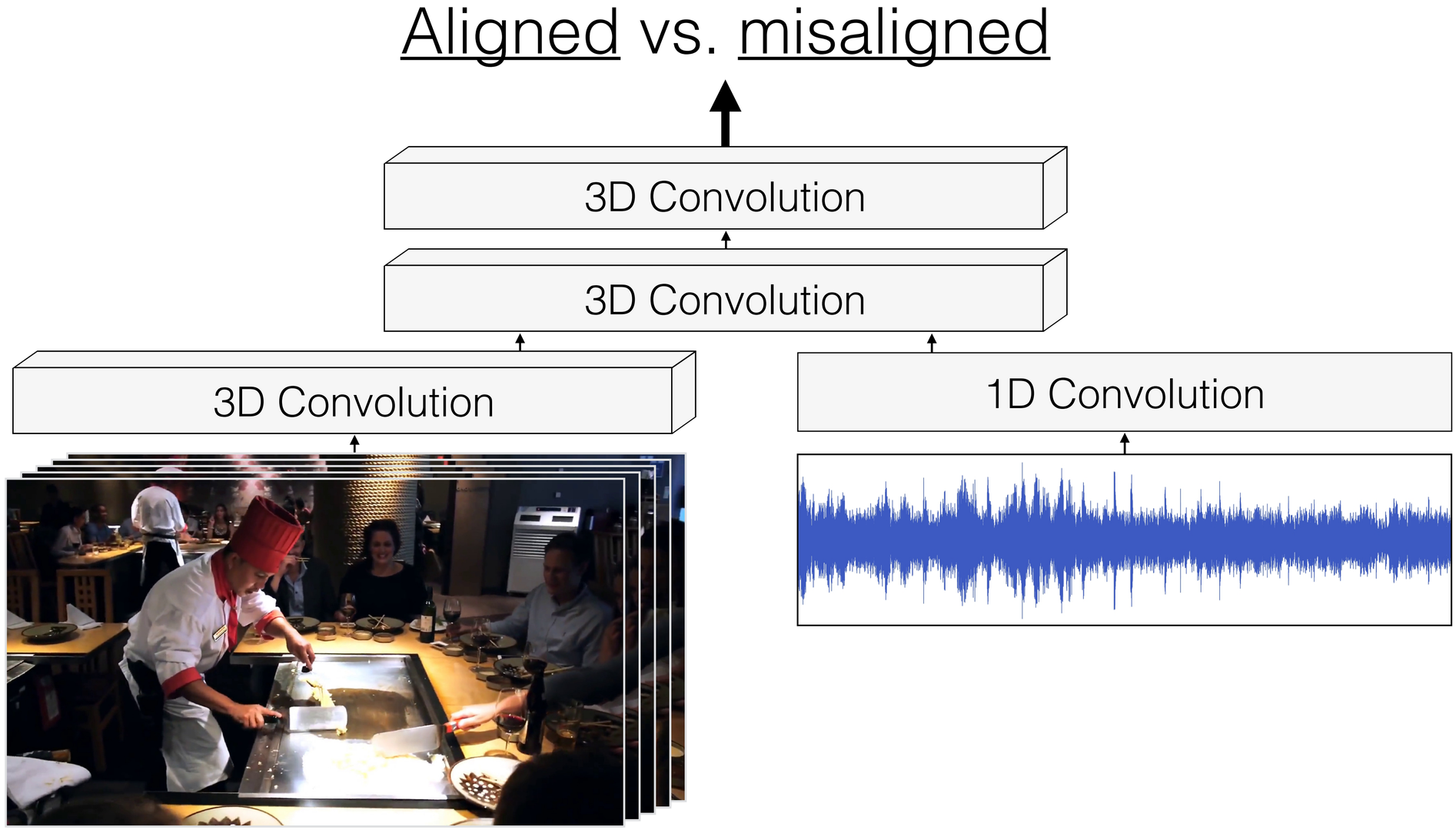
Idea #2: time-shifted pairs



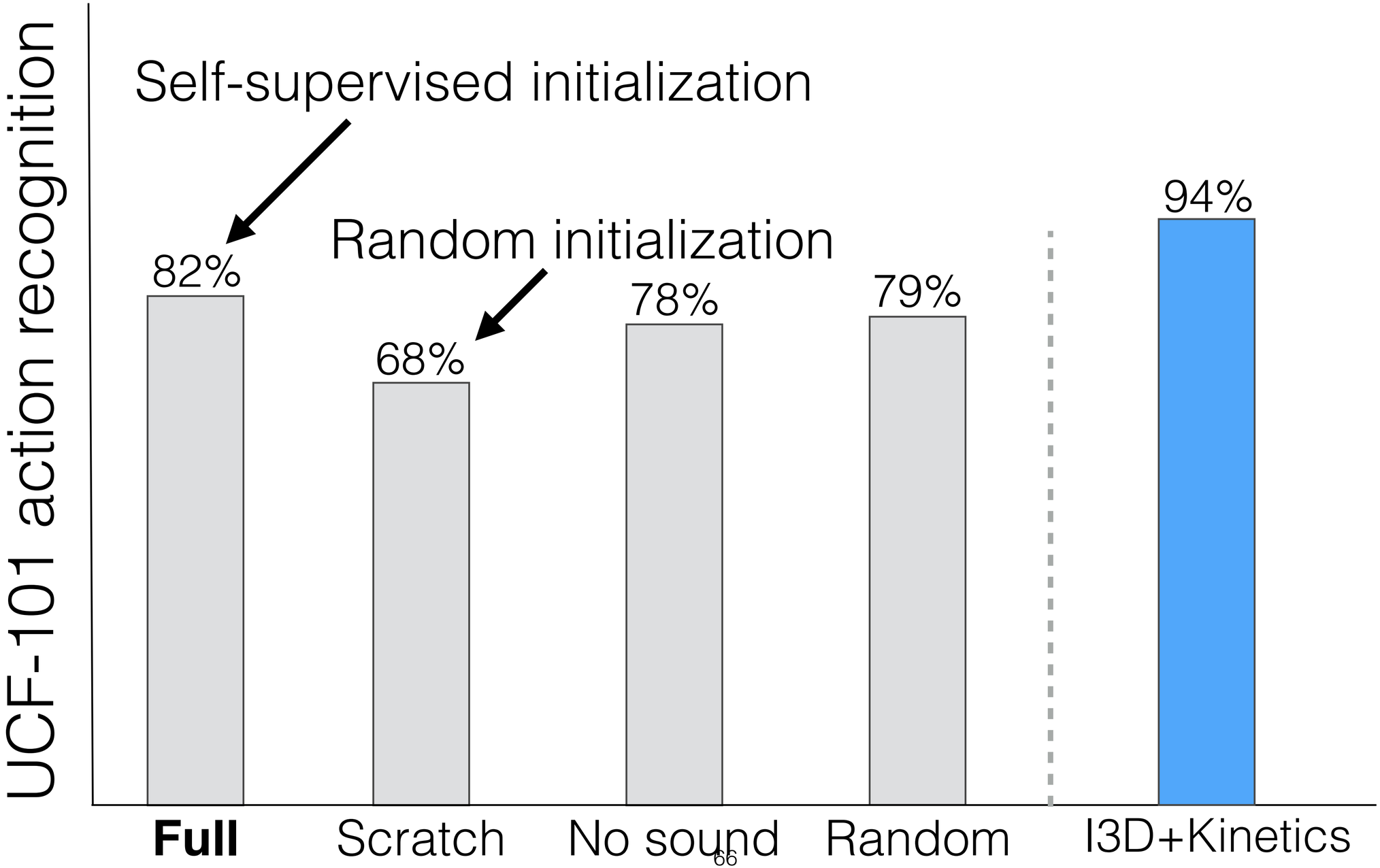
Idea #2: time-shifted pairs



Learning an audio-visual representation

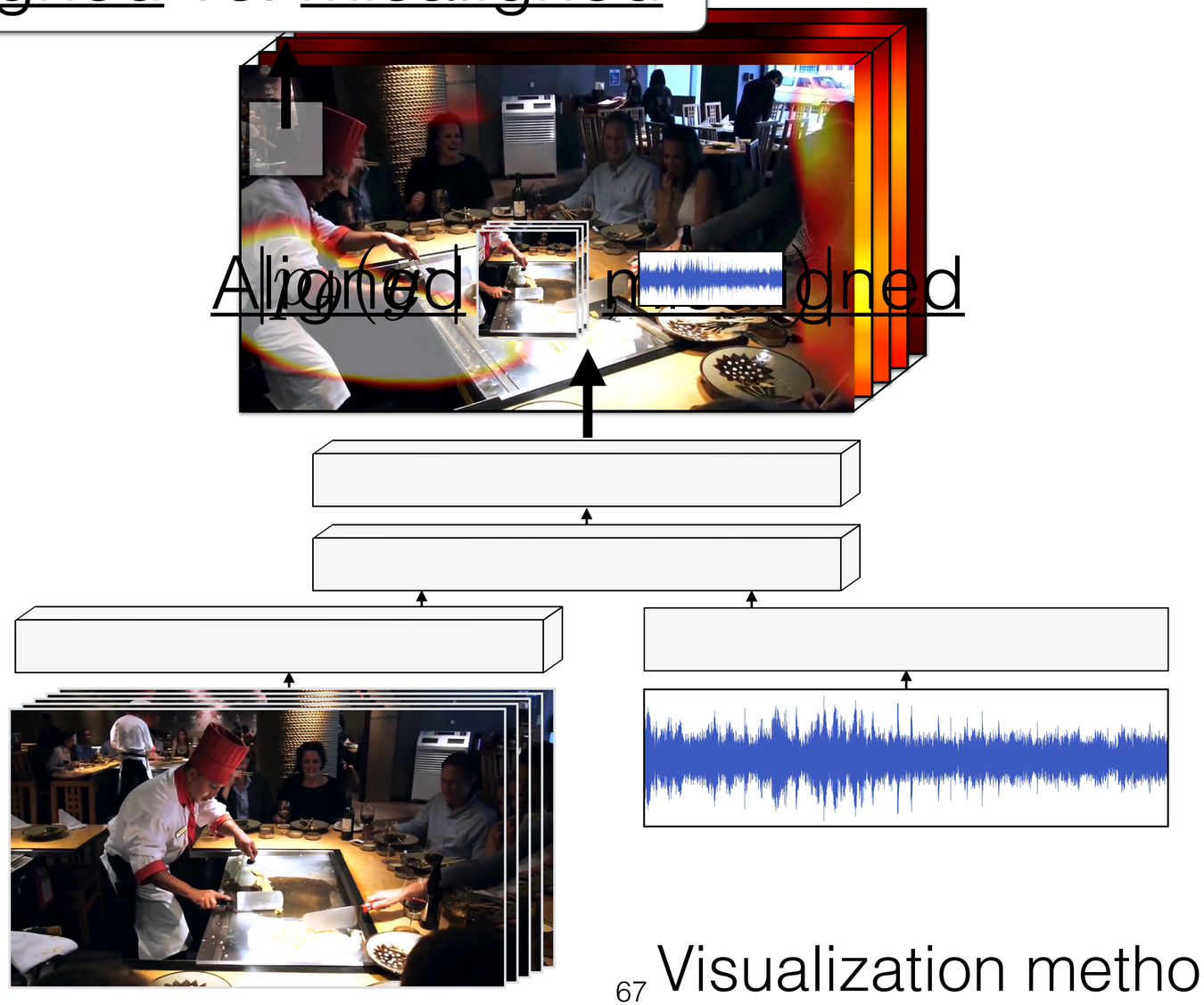


Action recognition



What did the model learn?

Aligned vs. misaligned



Visualization method from (Zhou 2016)

Dribbling basketball

Dribbling basketball

Dribbling basketball

Playing organ

Playing organ

Playing organ

Chopping wood

Chopping wood

Chopping wood

Application: on/off-screen source separation

Cocktail party problem

Underdetermined!

a	=	b	+	c
Sound mixture		On-screen		Off-screen

Cocktail party problem

- Independent component analysis (ICA) requires multiple microphones
- Generative models (Roweis 2001) are hard to learn

$p(\mathrm{b})$ $p(\mathrm{c})$

a = b + c

Sound mixture On-screen Off-screen

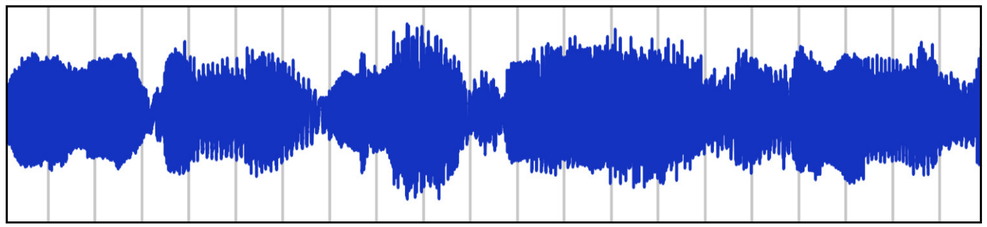
Cocktail party problem

- Independent component analysis (ICA) requires multiple microphones
- Generative models (Roweis 2001) are hard to learn
- Discriminative prediction (Hershey 2016, …)

$$p(\text{b, c} \mid \text{a})$$

a	=	b	+	c
Sound mixture		On-screen		Off-screen

Learning from synthetic sound mixtures

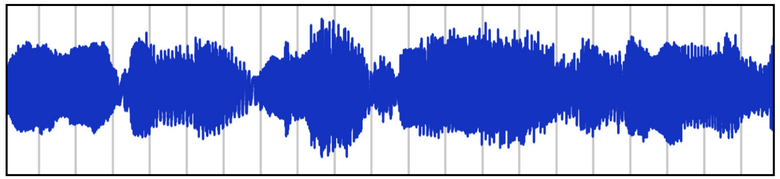


Synthetic mixture

Offscreen

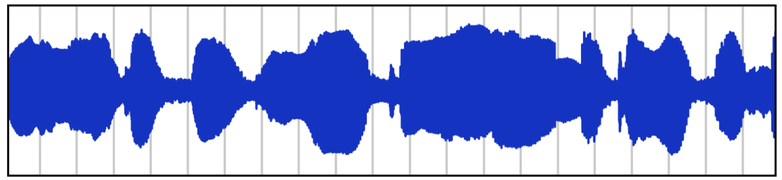
+

Learning from synthetic sound mixtures

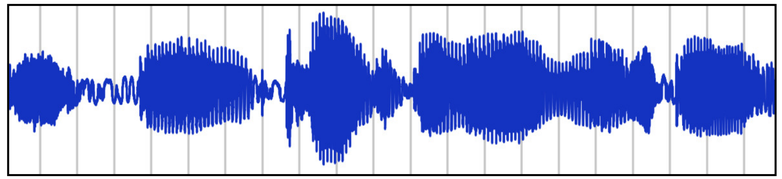


Synthetic mixture

Regression



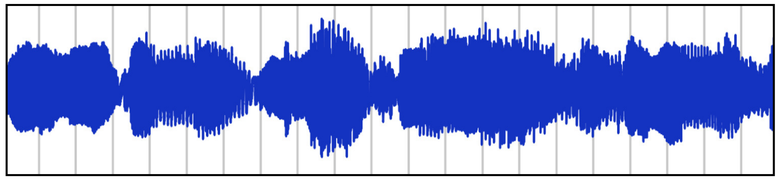
On-screen



Off-screen

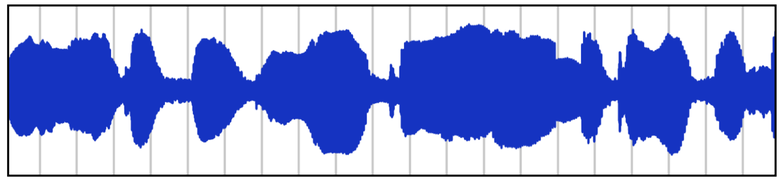
Related work: audio-only separation (Hershey 2016)

Learning from synthetic sound mixtures

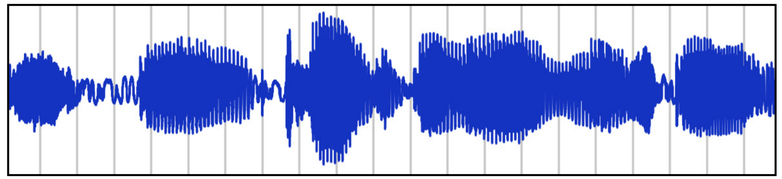


Synthetic mixture

Regression



On-screen

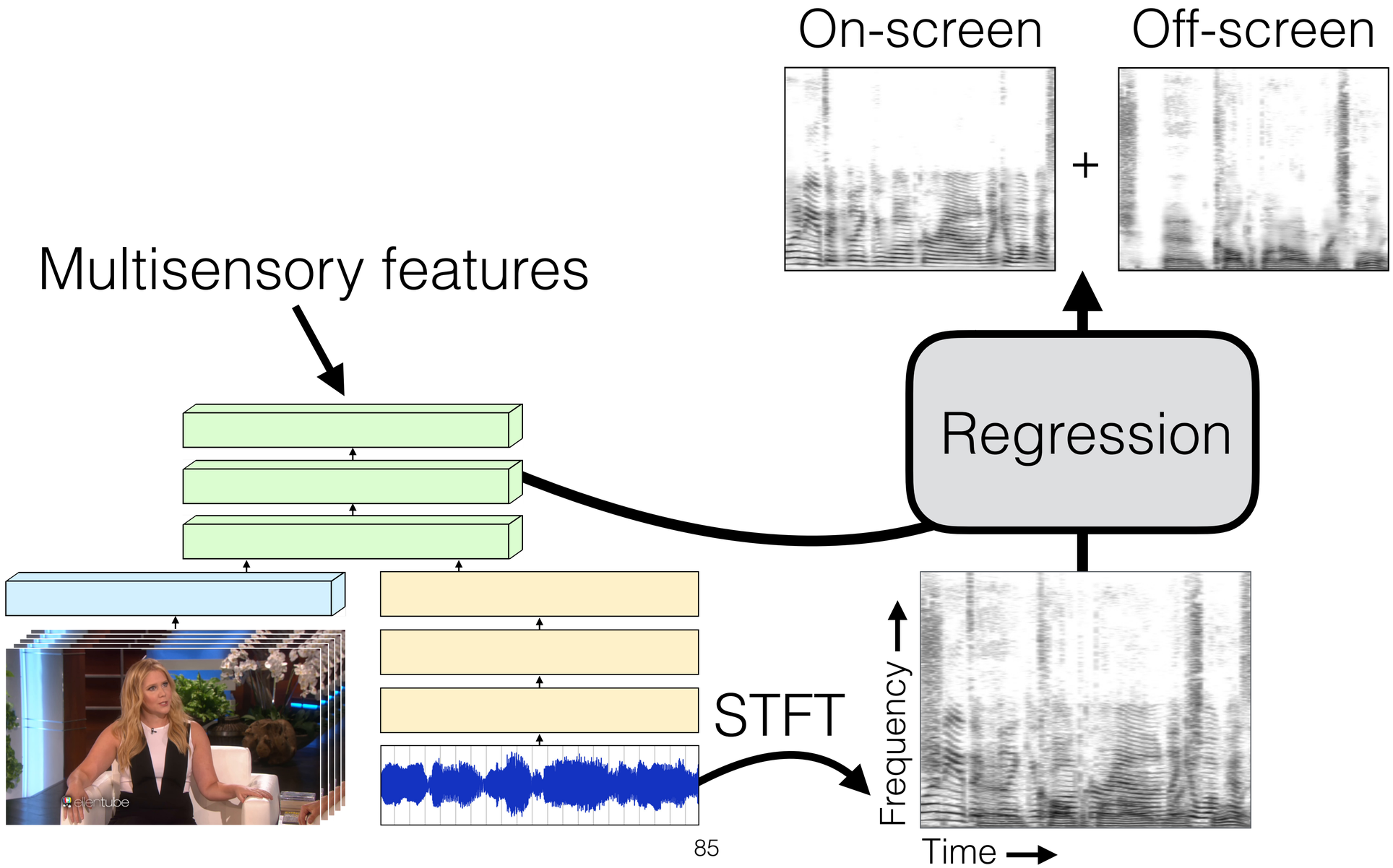


Off-screen

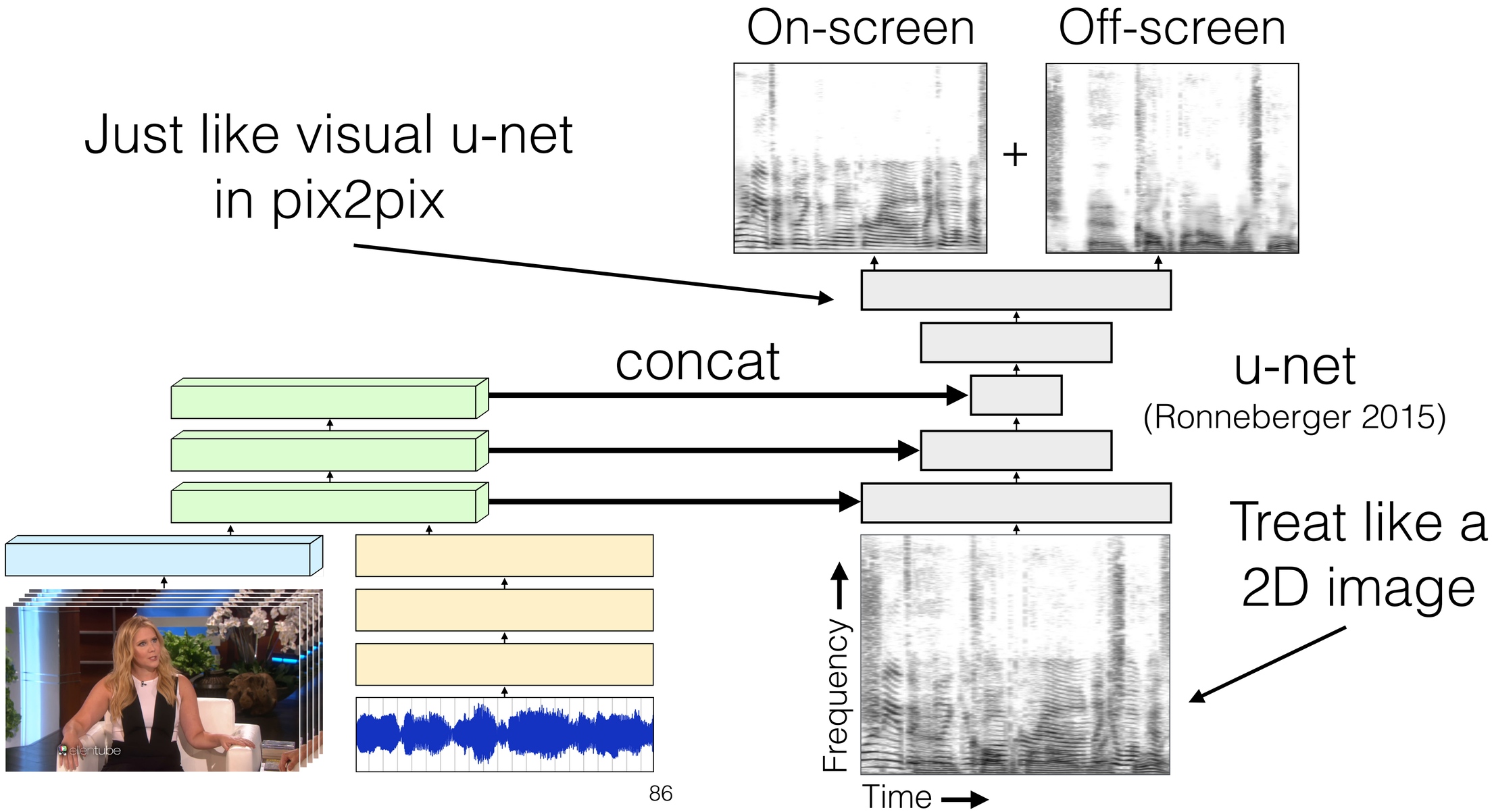
Audio separation: e.g. (Hershey 2016)

Concurrent audio-visual work: (Gao 2018, Afouras 2018, Gabbay 2018)

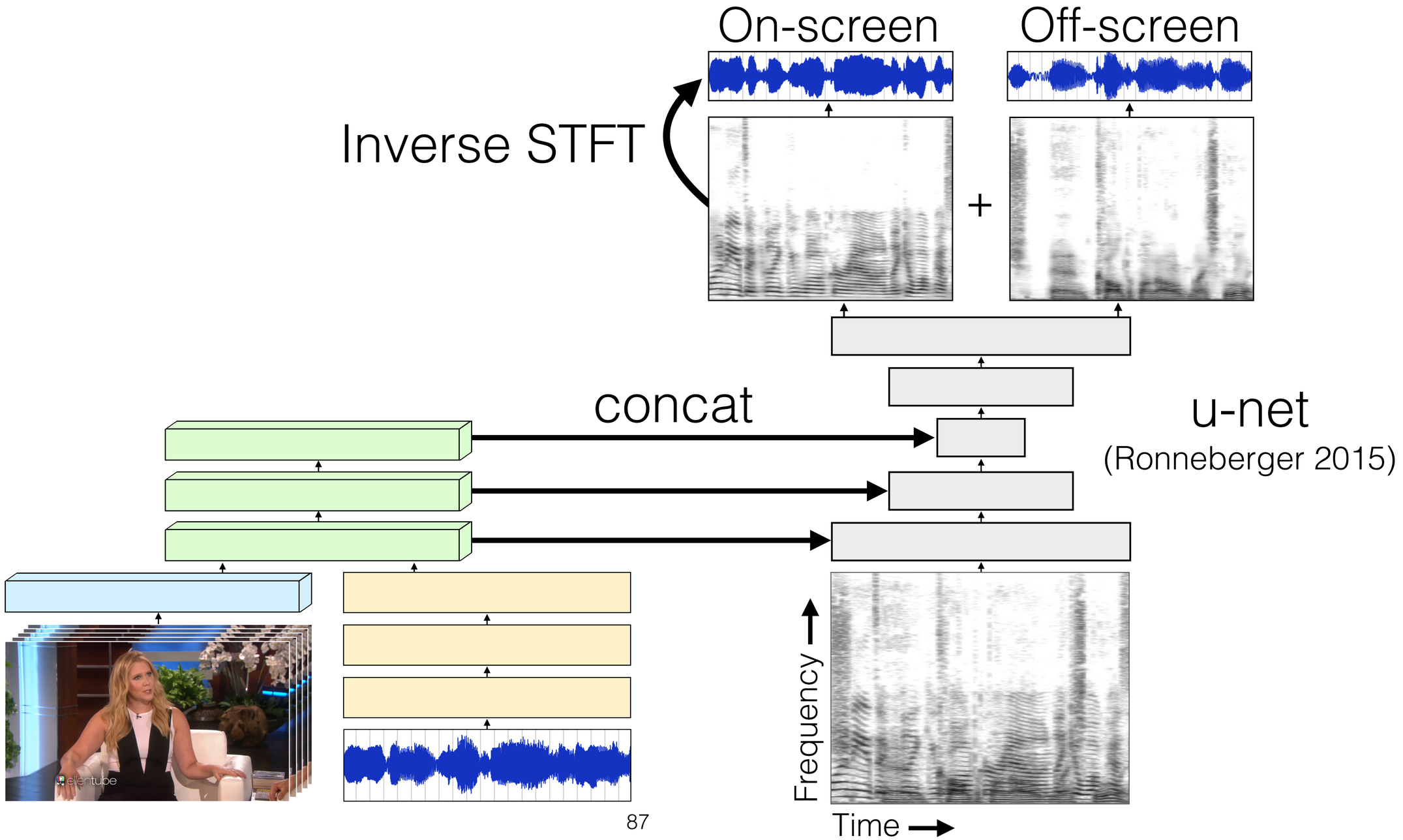
On/off-screen source separation



On/off-screen source separation



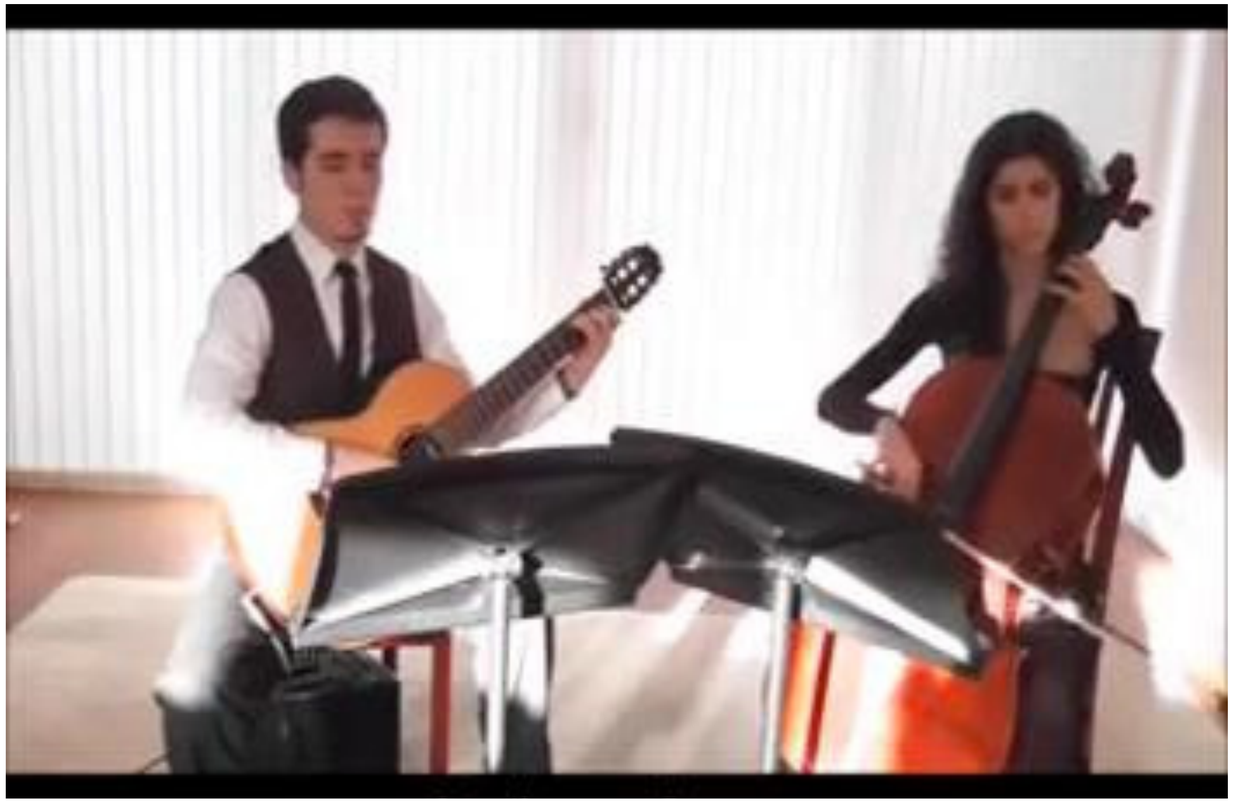
On/off-screen source separation



THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA VOLUME 25, NUMBER 5 SEPTEMBER, 1953

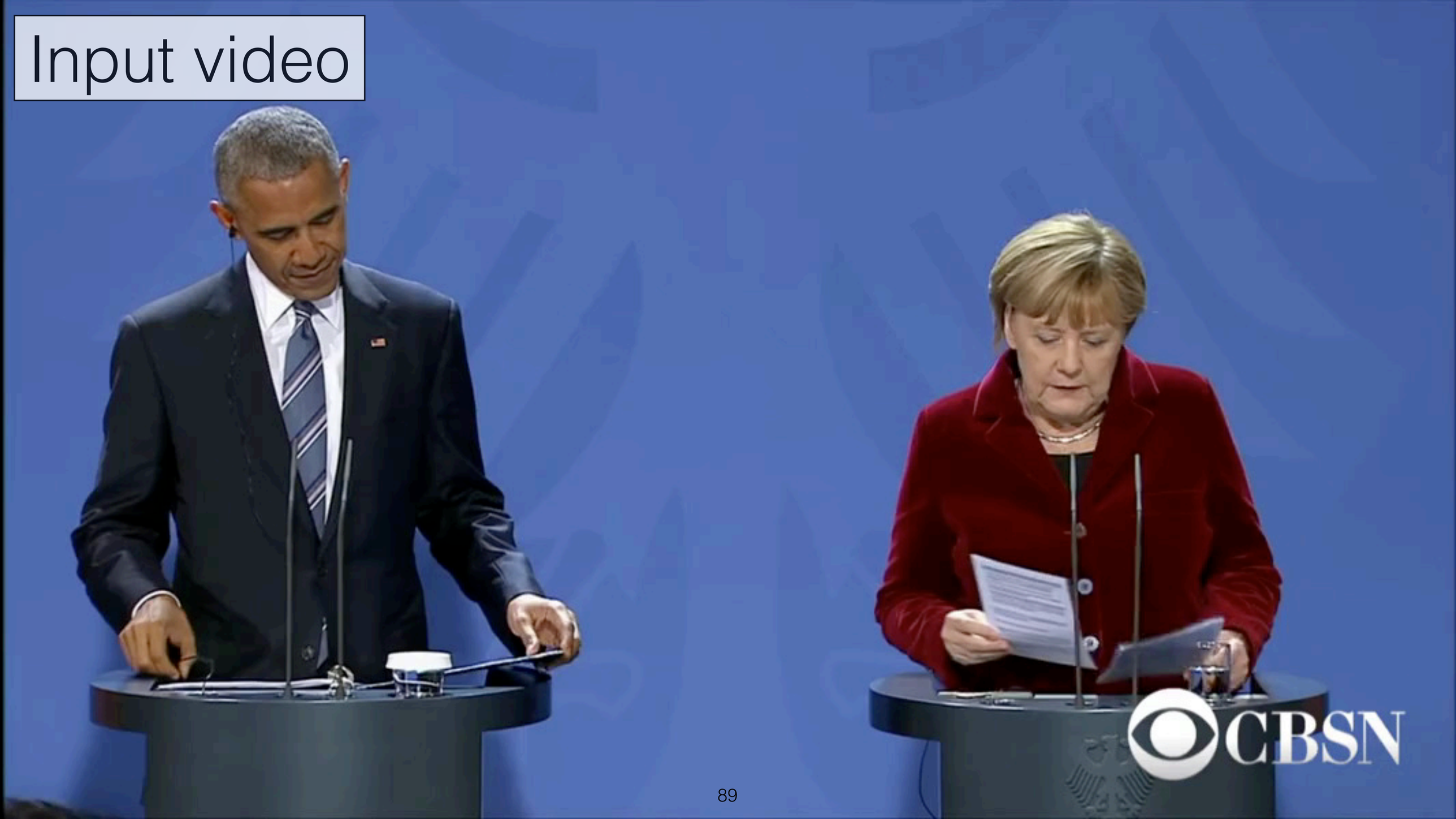
Some Experiments on the Recognition of Speech, with One and with Two Ears*

E. COLIN CHERRY
Imperial College, University of London, England, and Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts

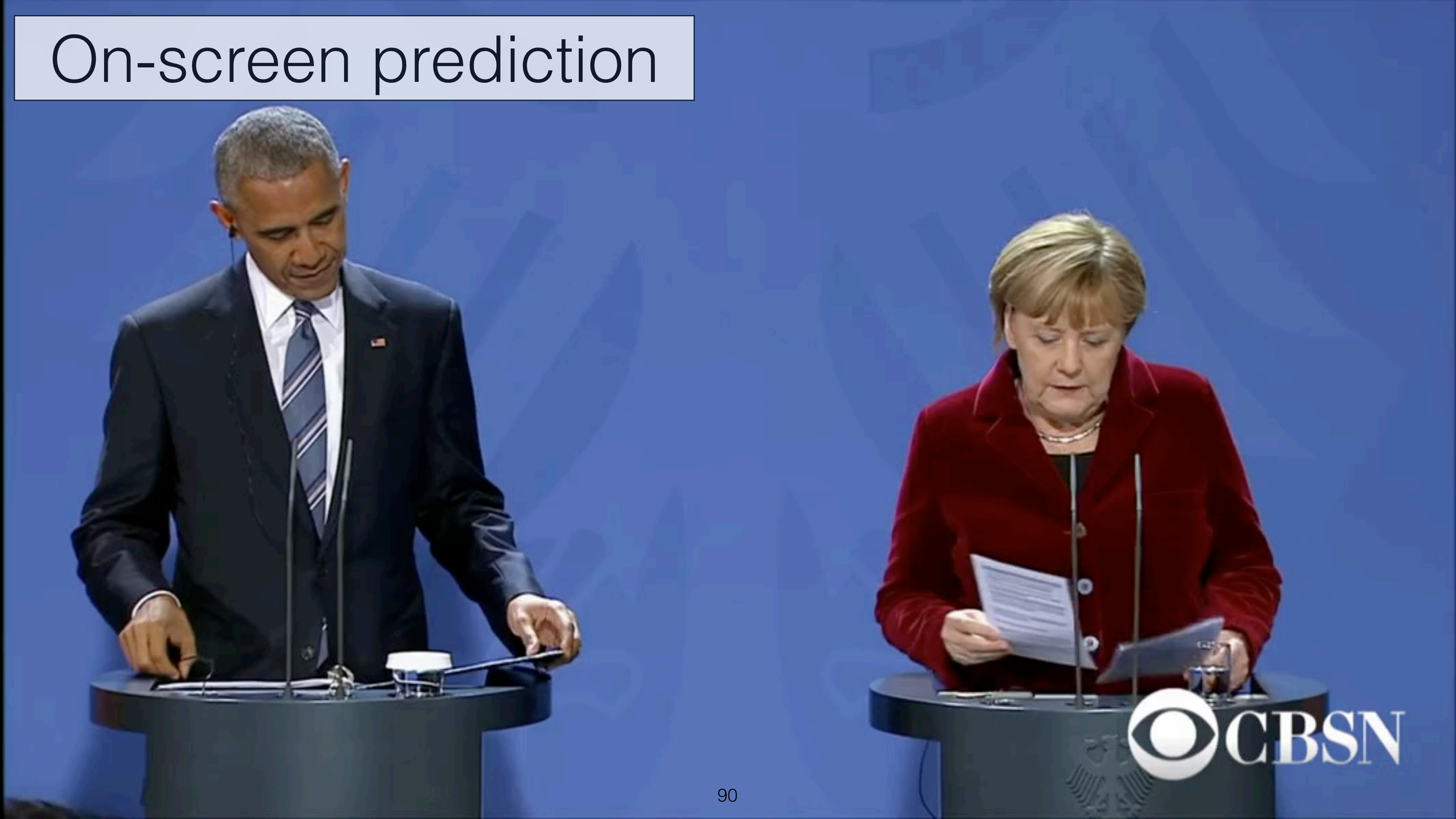


The cocktail party problem

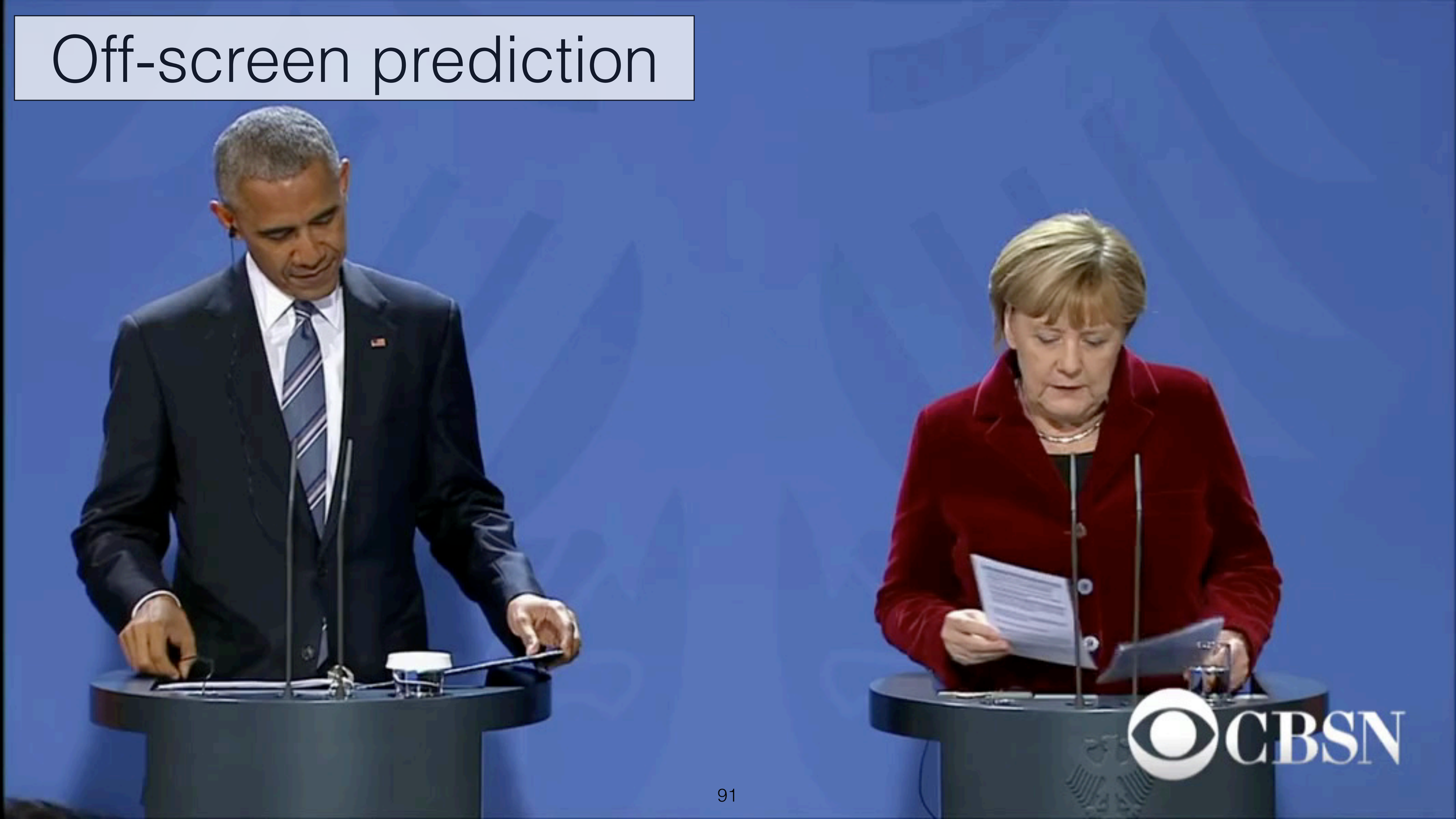
Source: Torralba, Isola, Freeman


Input video
CBSN

On-screen prediction



Off-screen prediction



Input video


Las Vegas
ONE-ON-ONE
TRUMP CALLS FOR DOJ INVESTIGATION OF NY TIMES OP-ED
LIVE
CNN
6:04 PM PT
CUOMO PRIME TIME

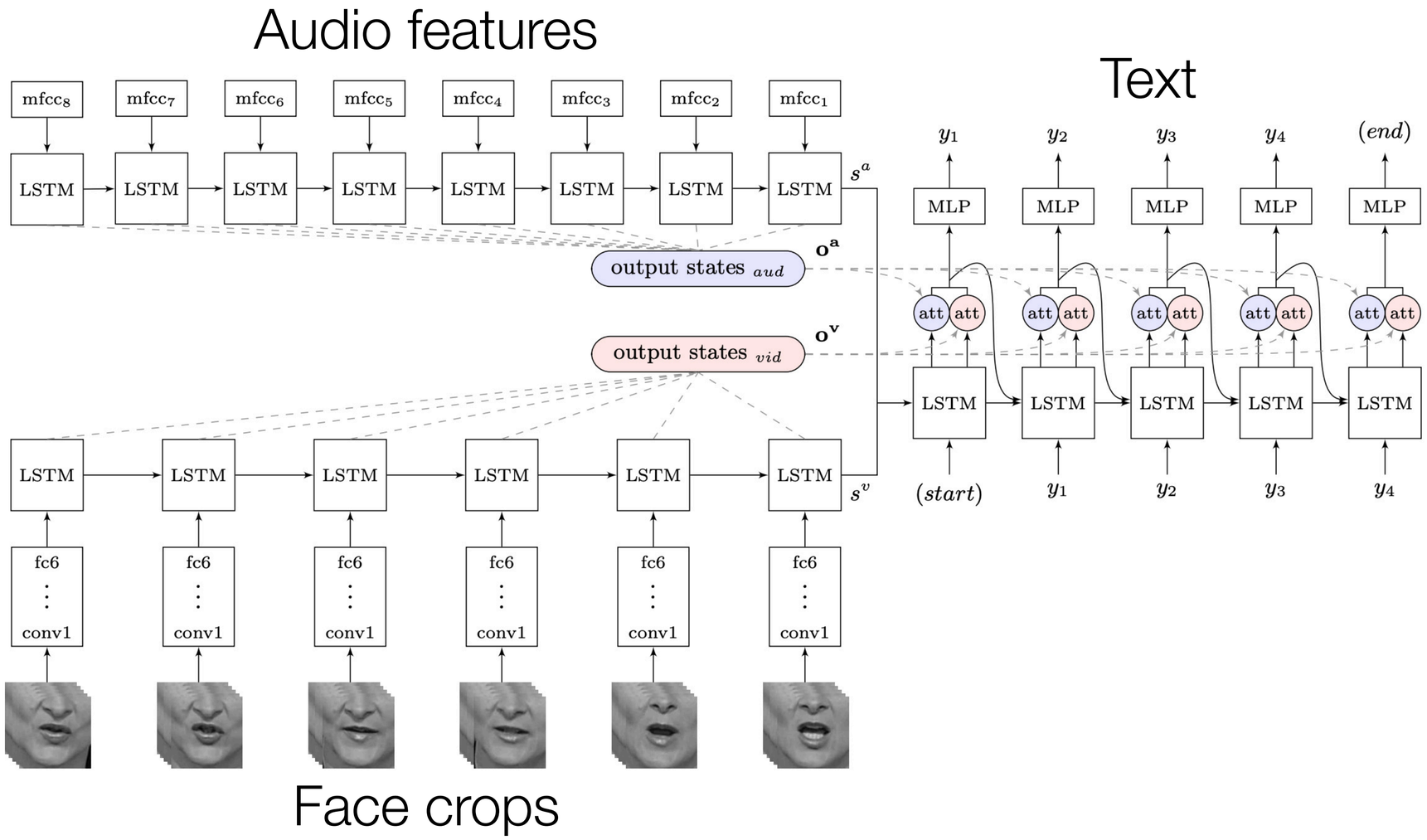
On-screen prediction

On-screen prediction

Looking to Listen at the Cocktail Party:
A Speaker-Independent Audio-Visual Model for Speech Separation

Ariel Ephrat Inbar Mosseri Oran Lang Tali Dekel Kevin Wilson
Avinatan Hassidim William T. Freeman Michael Rubinstein

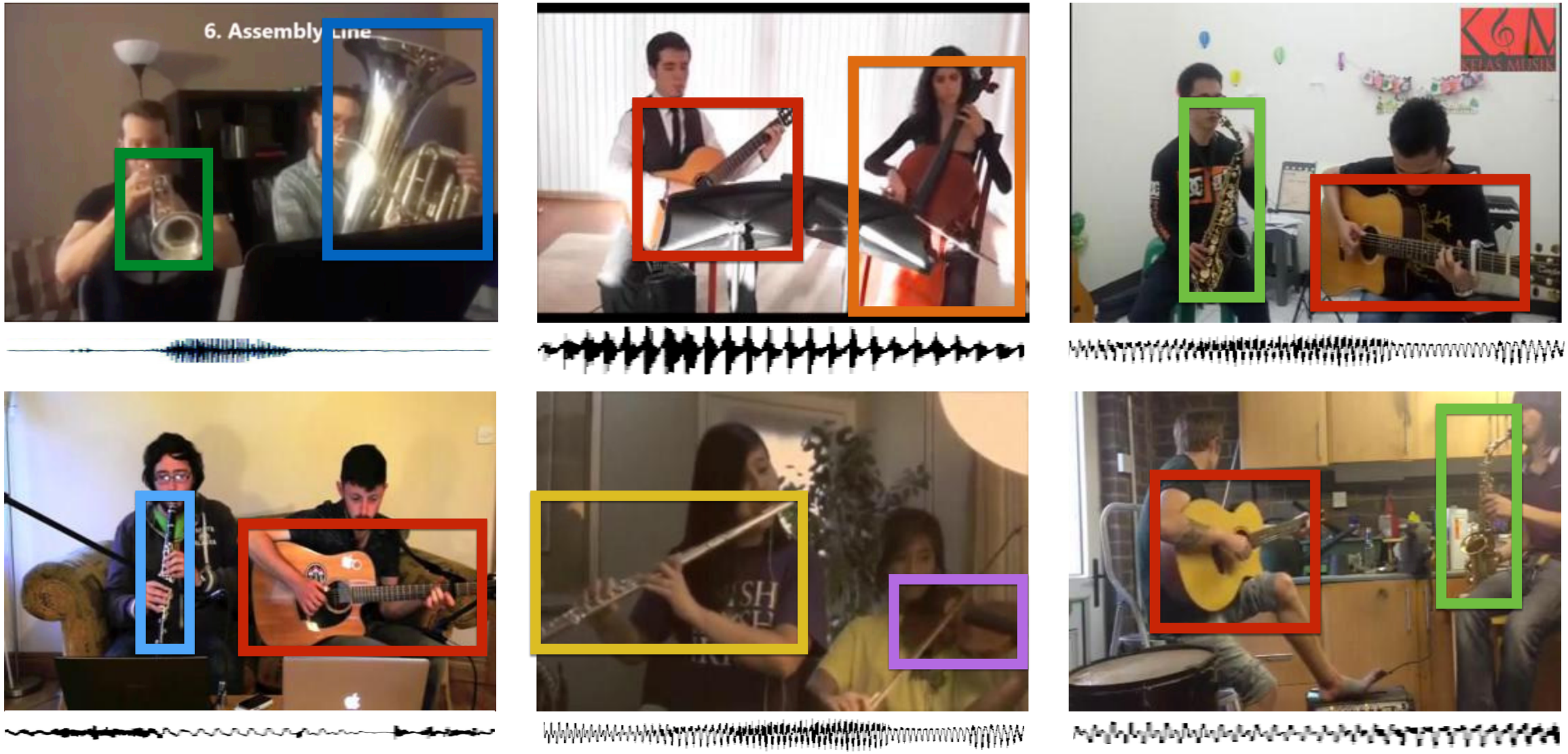
Lip reading



[Chung et al., "Lip reading sentences in the wild", 2017]

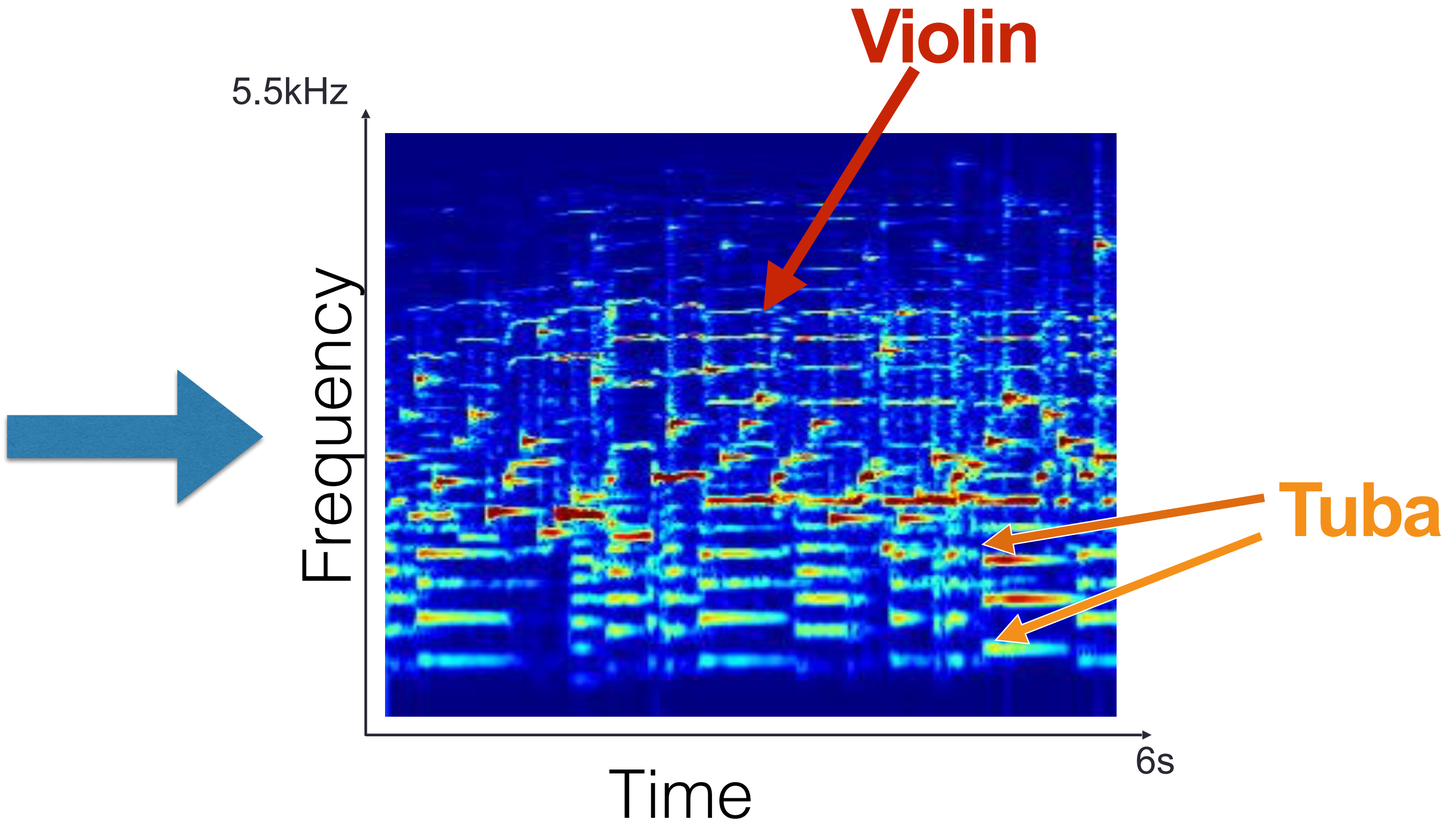
The video-to-text alignment is also generated by the model.

Discovering musical instruments

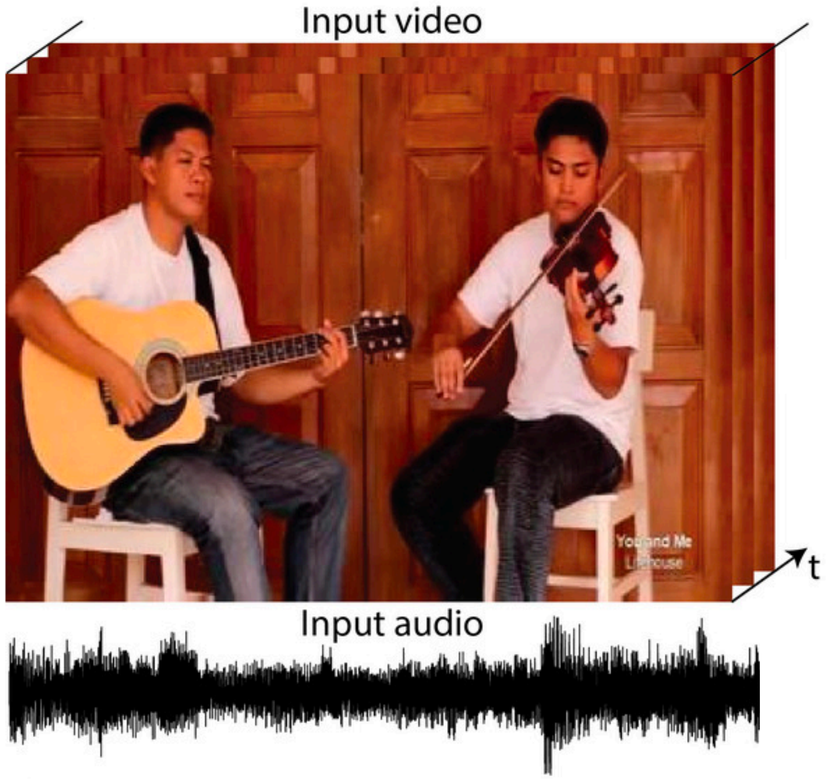


[Zhao et al., “The Sound of Pixels”, 2018]

Musical instrument separation

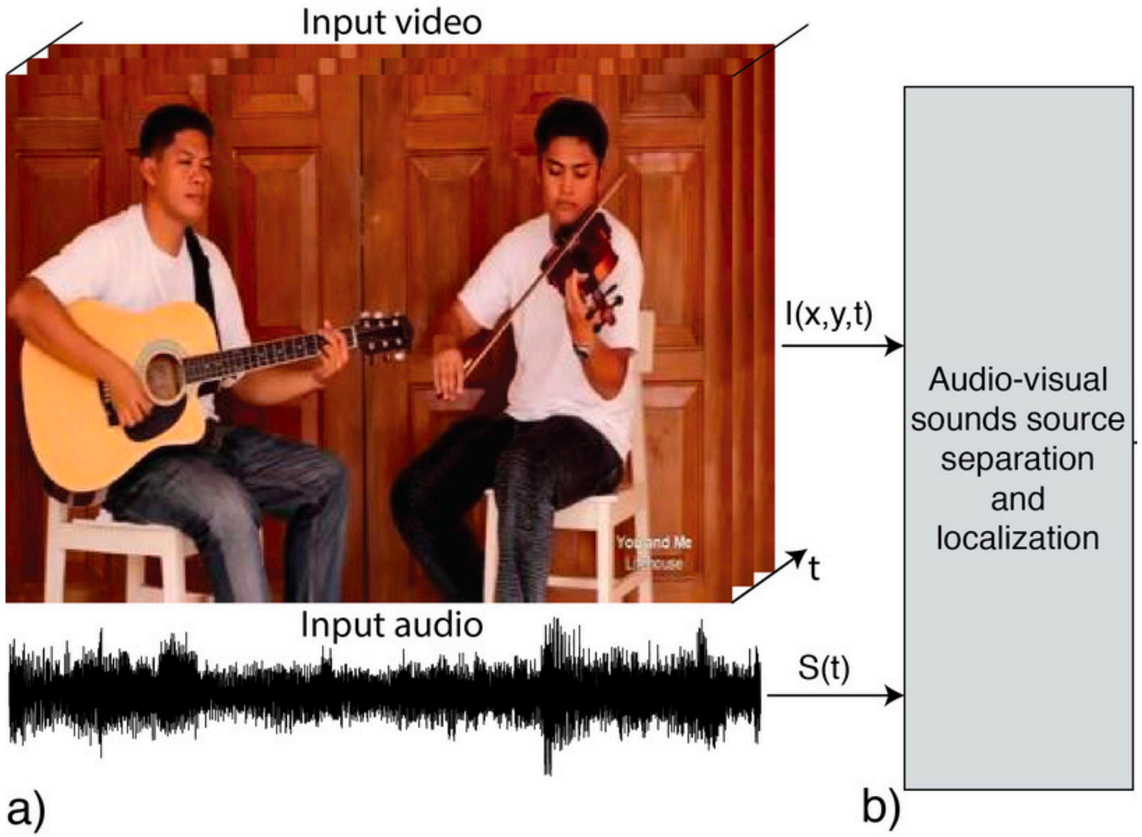


[Zhao et al., “The Sound of Pixels”, 2018]

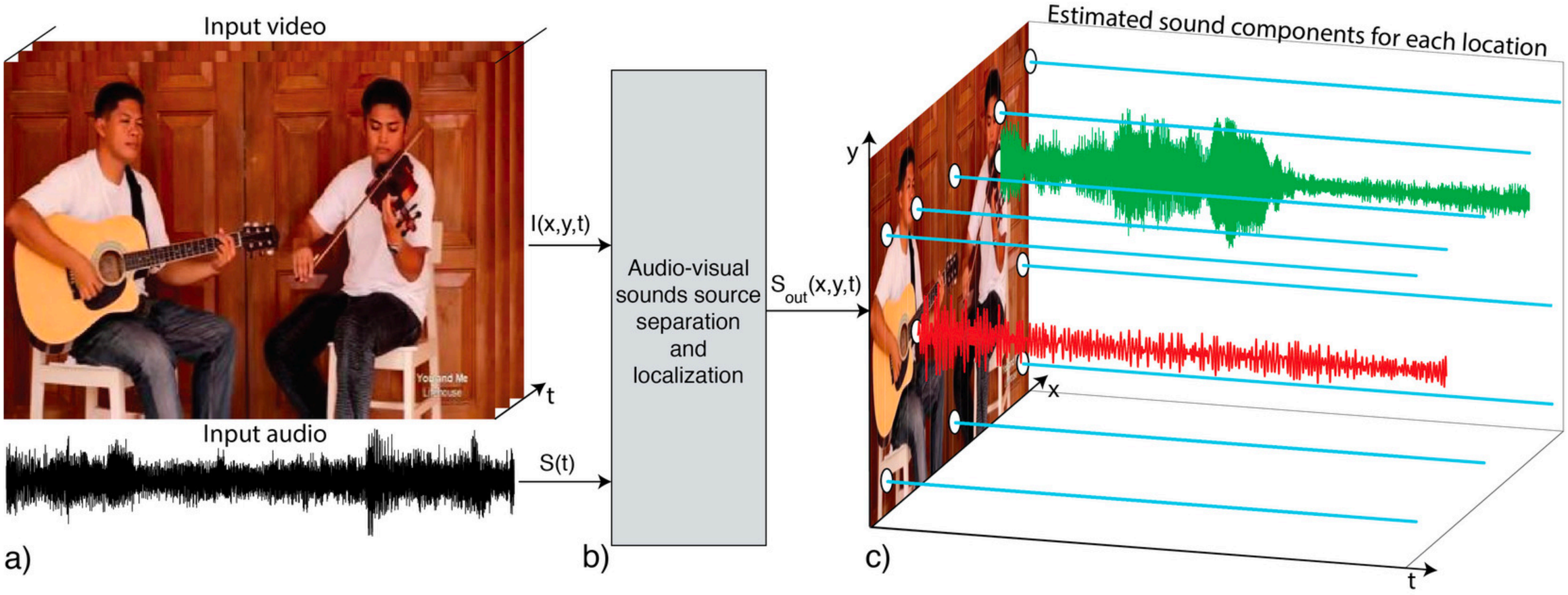


http://sound-of-pixels.csail.mit.edu/

[Zhao et al., “The Sound of Pixels”, 2018]

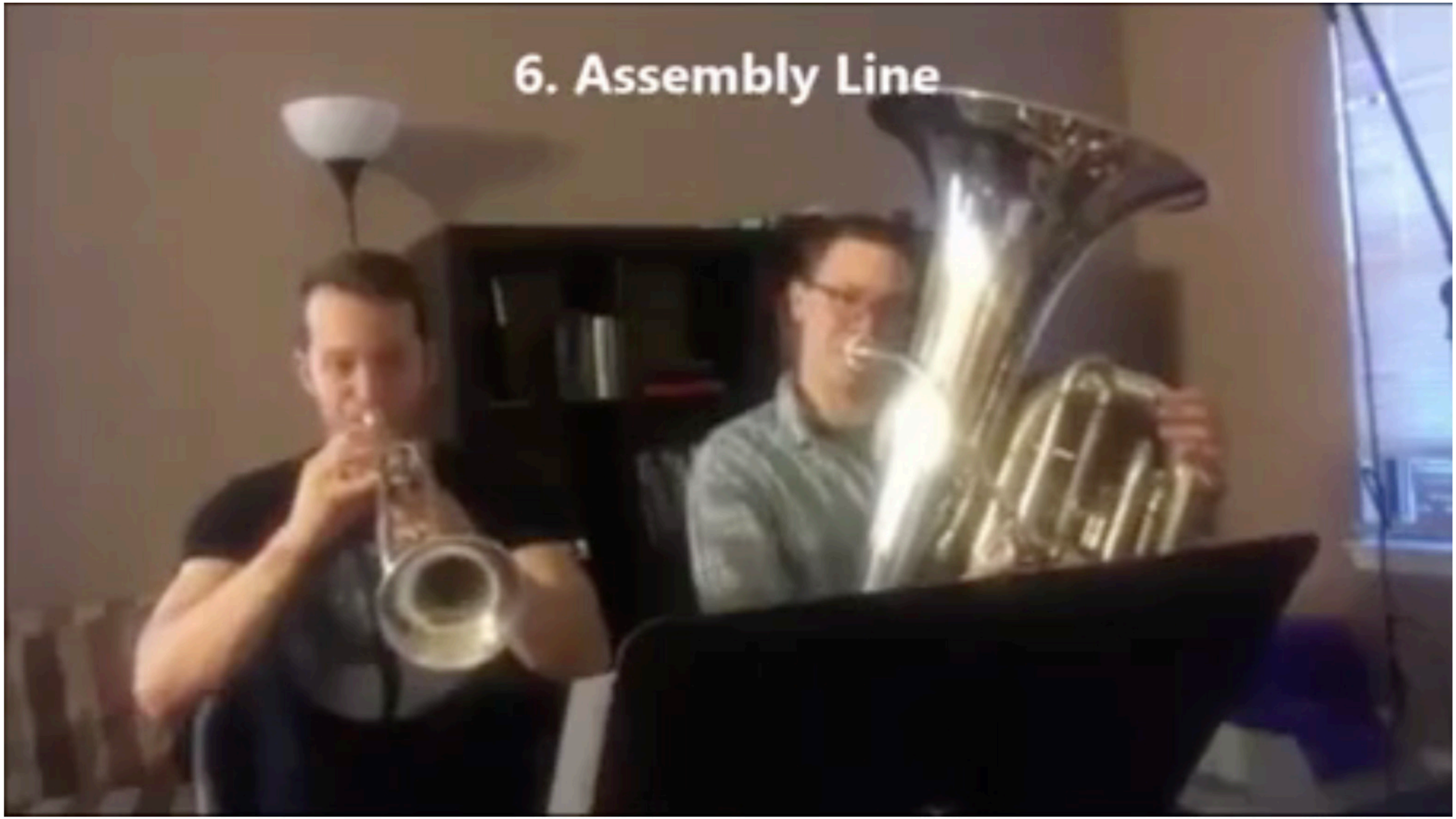


[Zhao et al., “The Sound of Pixels”, 2018]

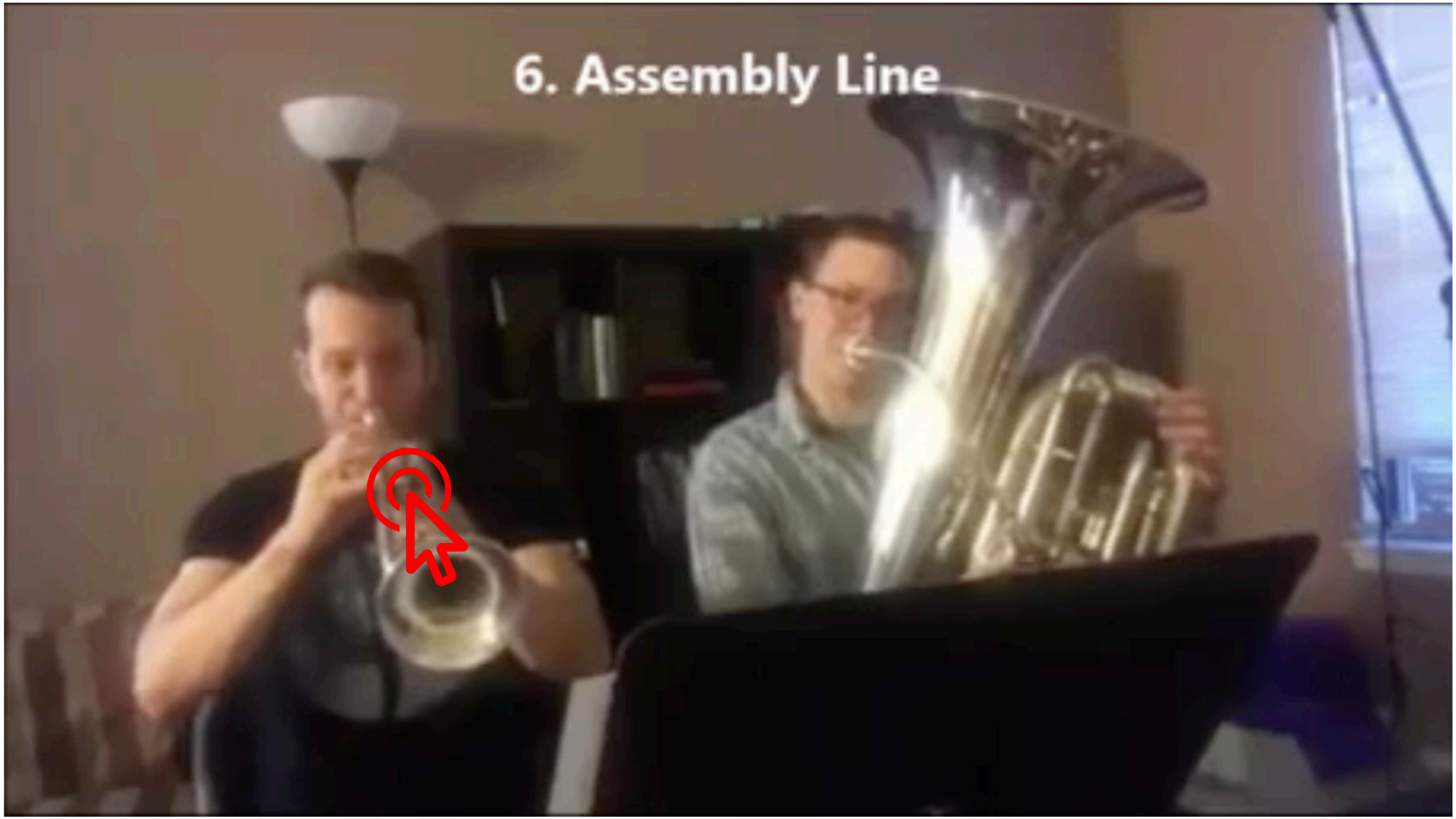


[Zhao et al., “The Sound of Pixels”, 2018]

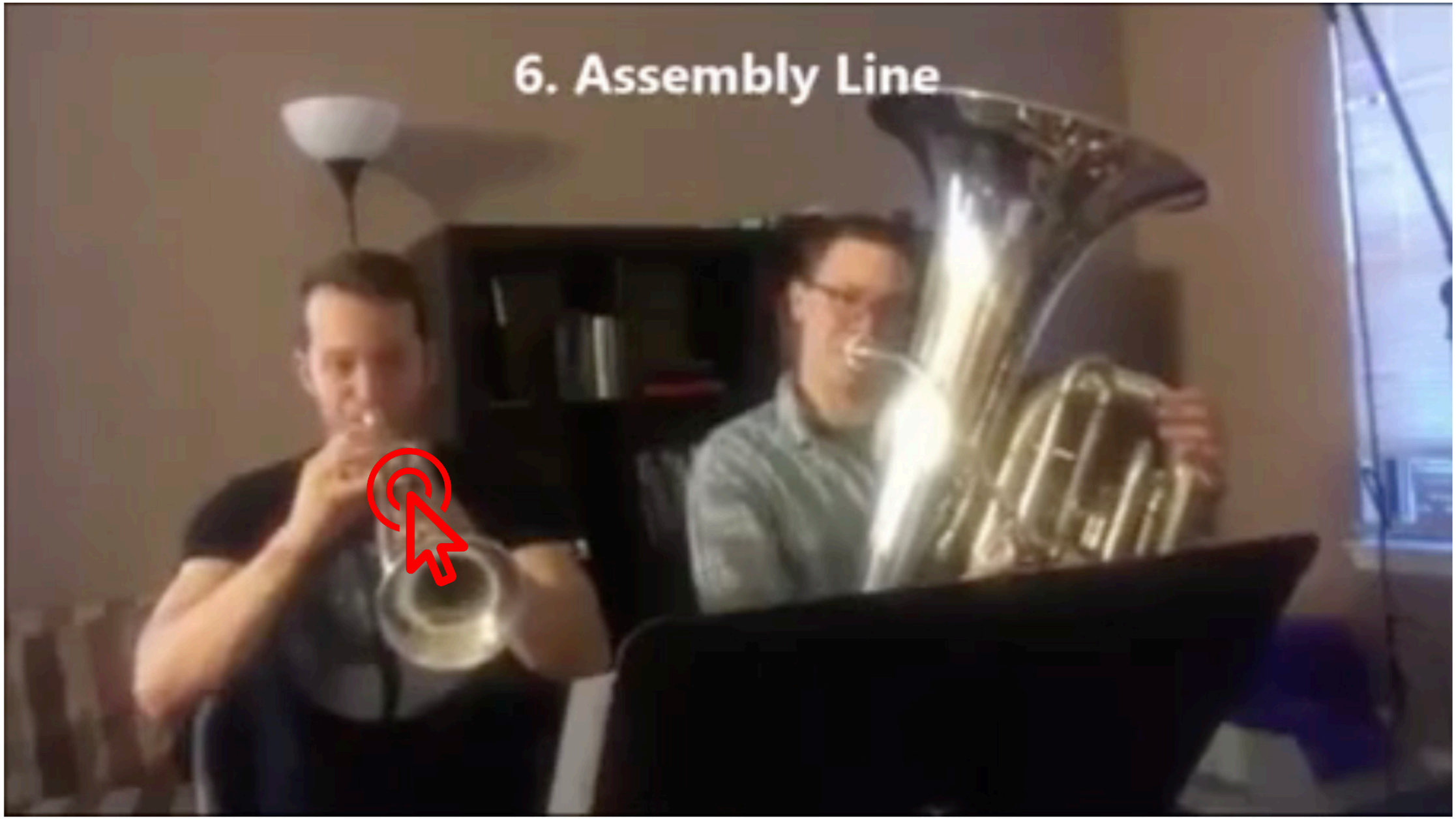
http://sound-of-pixels.csail.mit.edu/



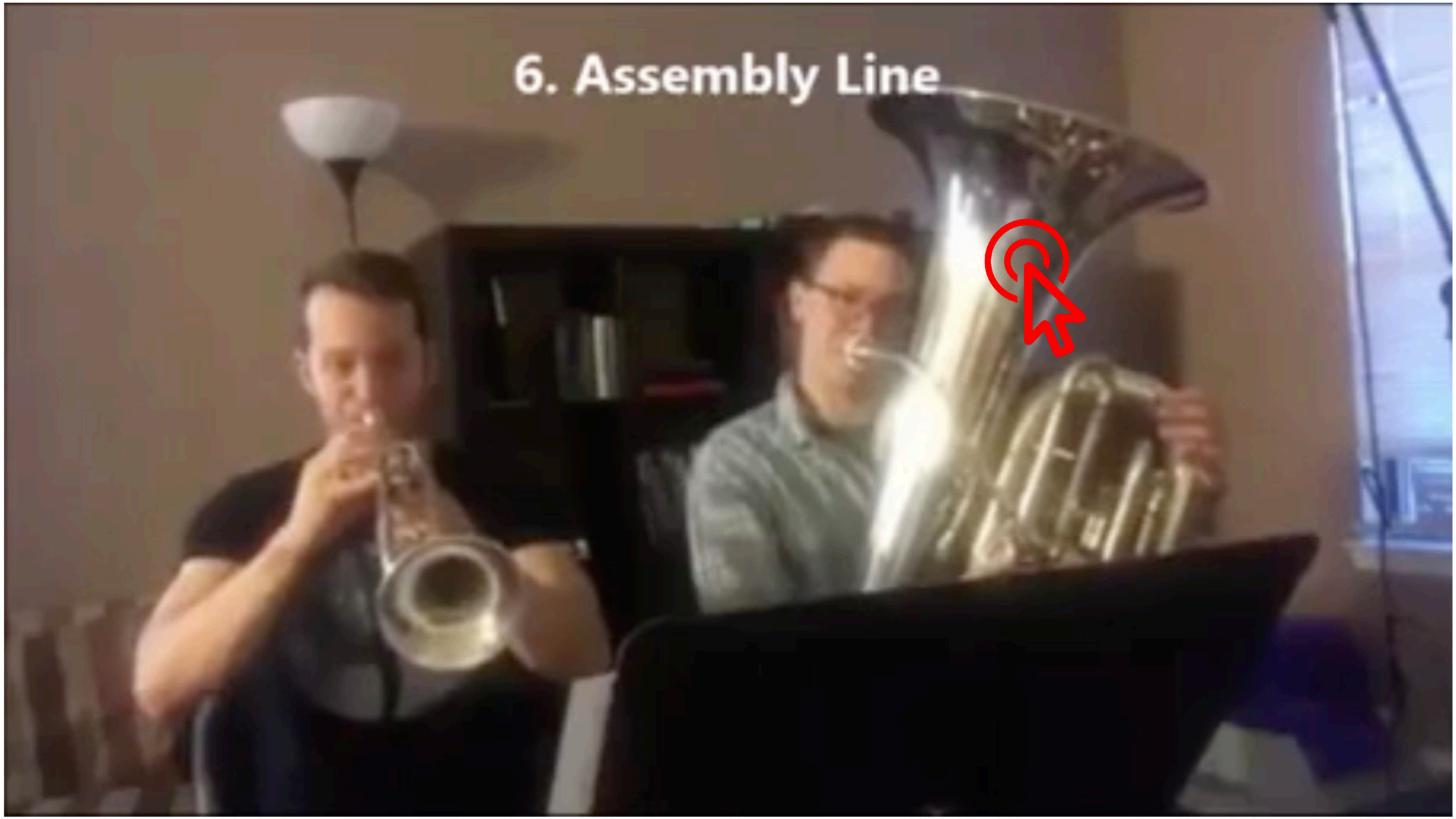
http://sound-of-pixels.csail.mit.edu/



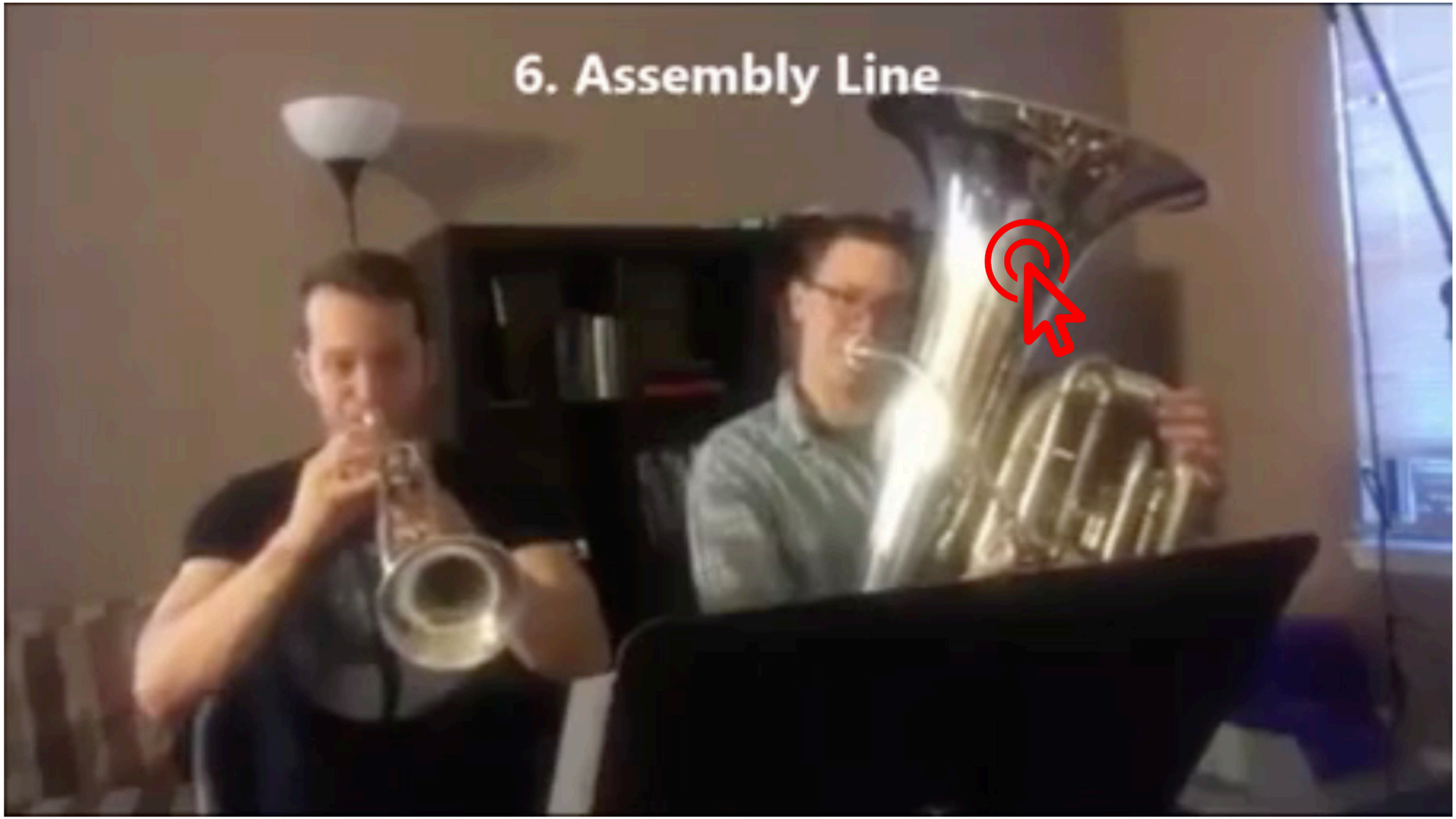
http://sound-of-pixels.csail.mit.edu/



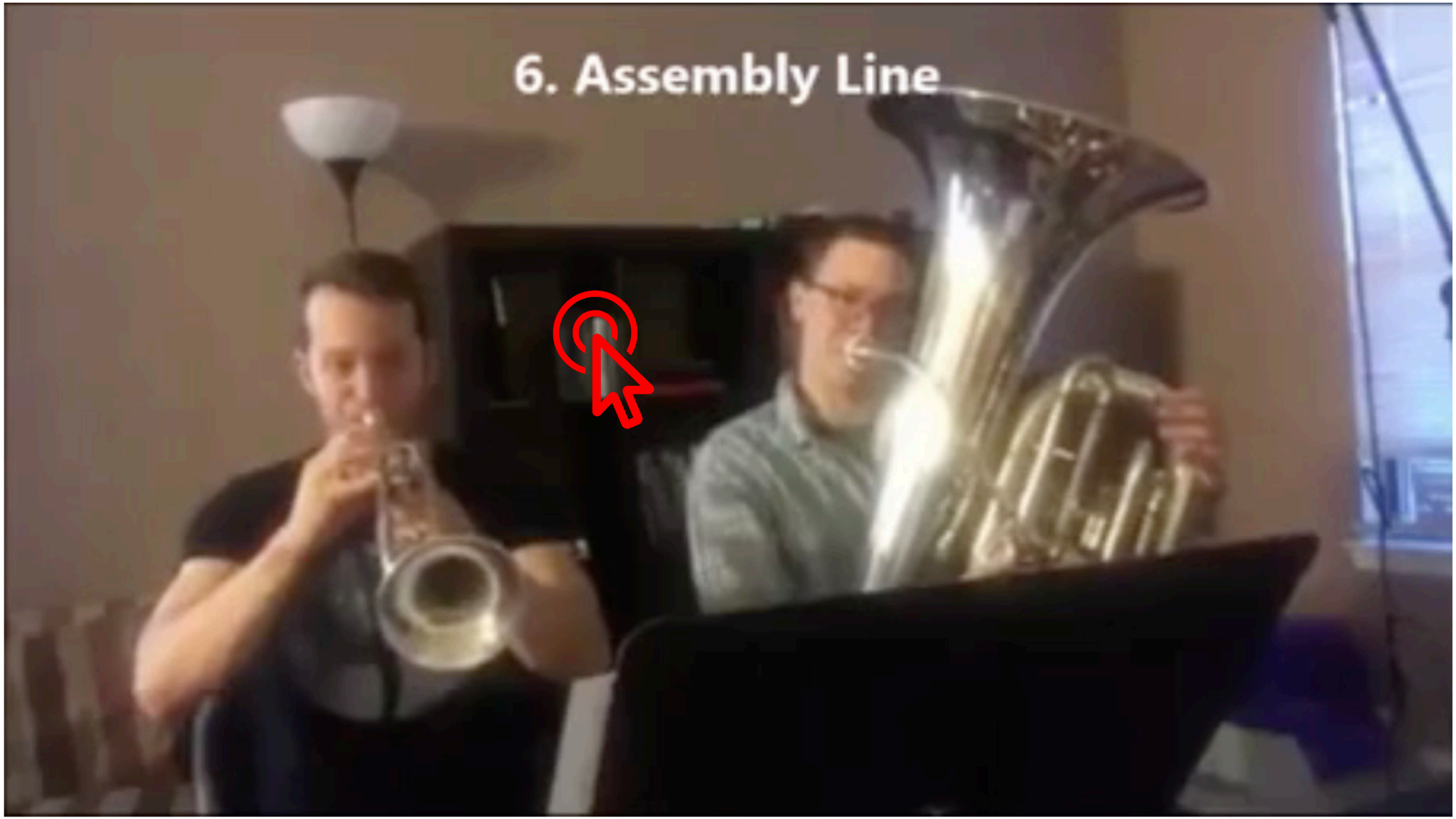
http://sound-of-pixels.csail.mit.edu/



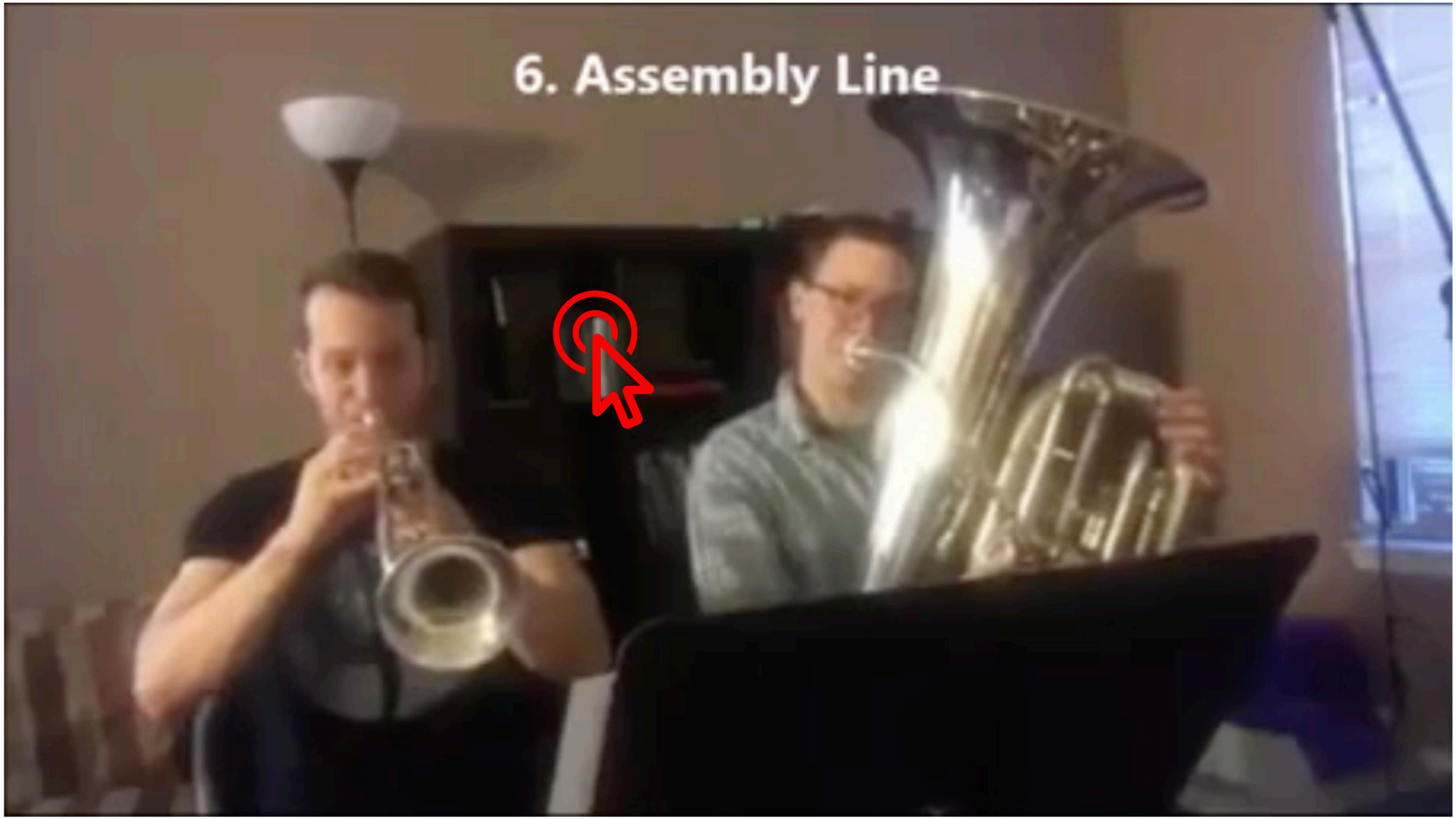
http://sound-of-pixels.csail.mit.edu/



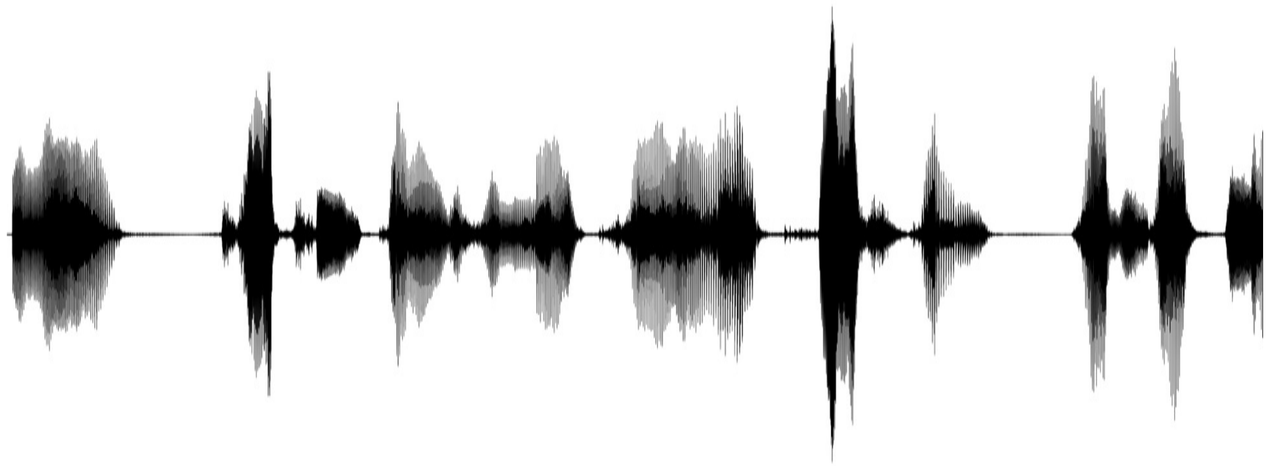
http://sound-of-pixels.csail.mit.edu/

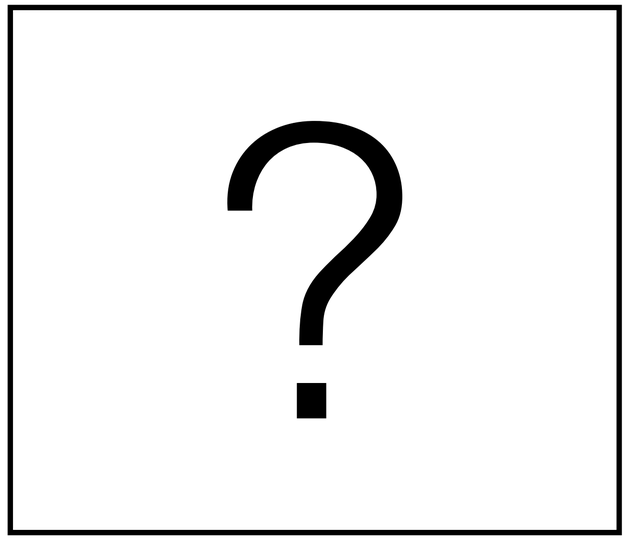


http://sound-of-pixels.csail.mit.edu/



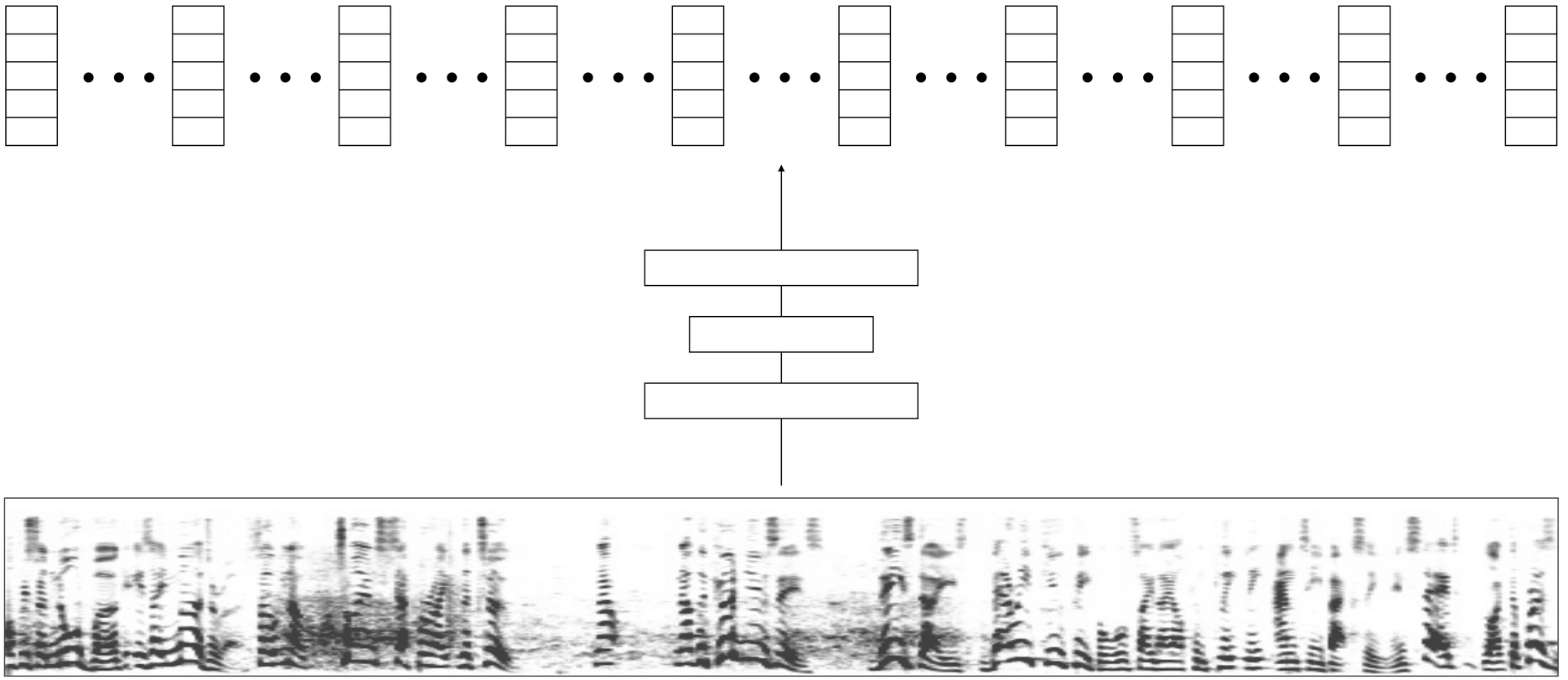
Predicting sight from sound



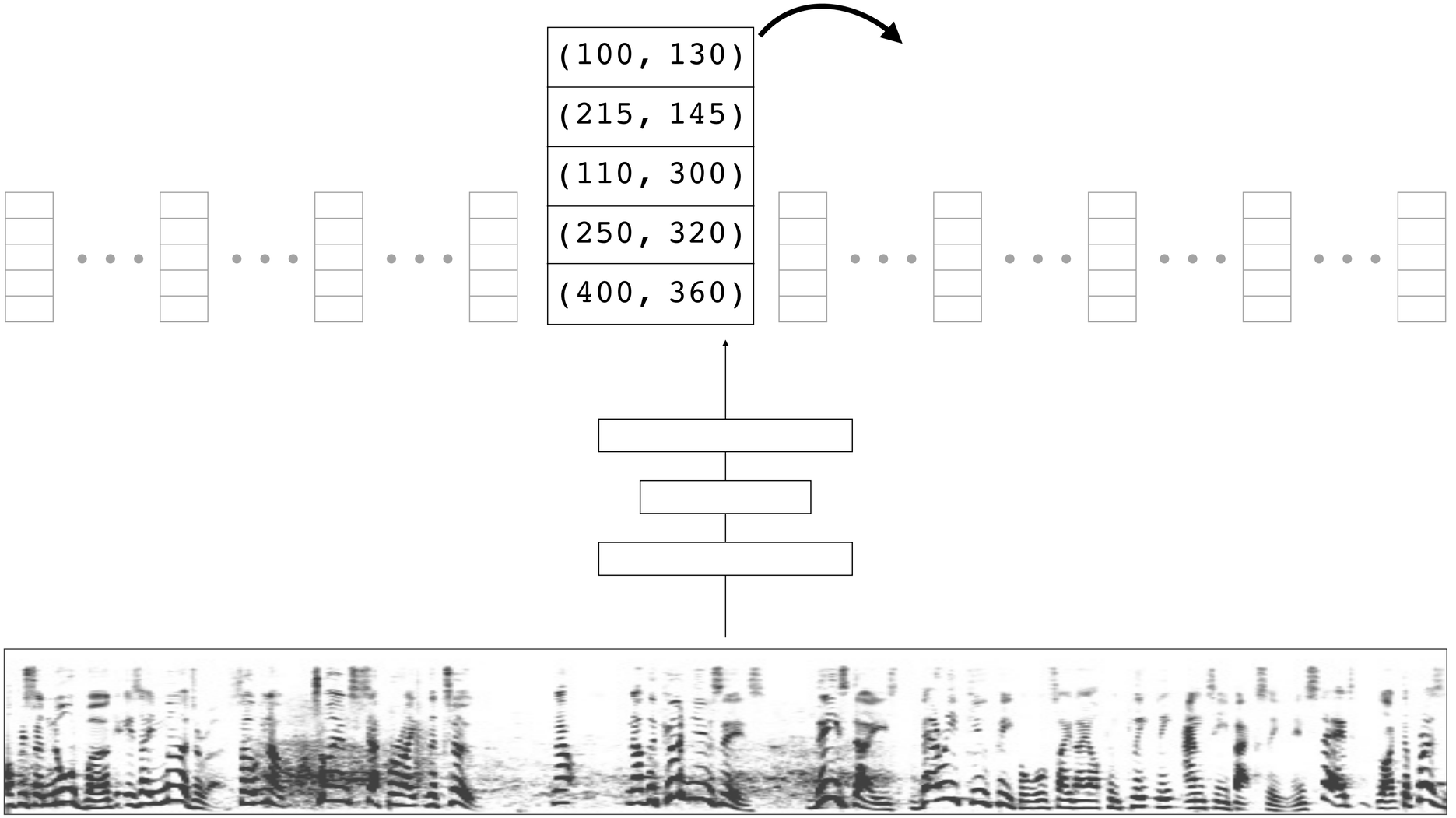



ellentube

Predicting gestures

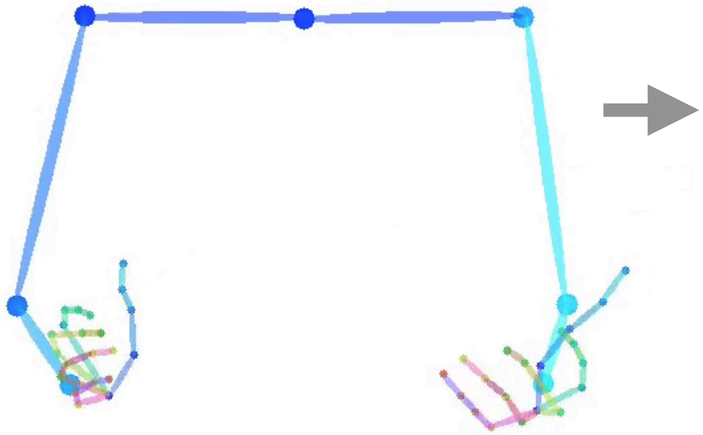


Predicting gestures



Synthesizing a video

Ground truth face



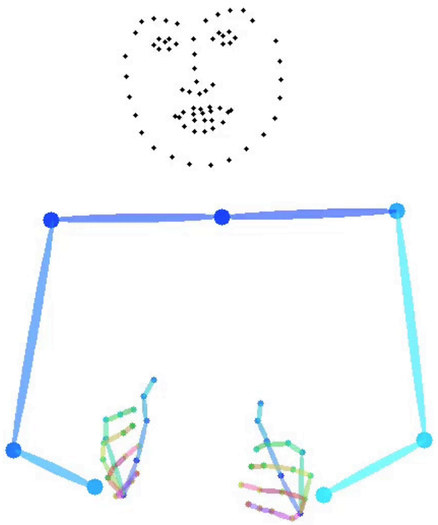
Audio input

Predicted gestures

Face is ground truth

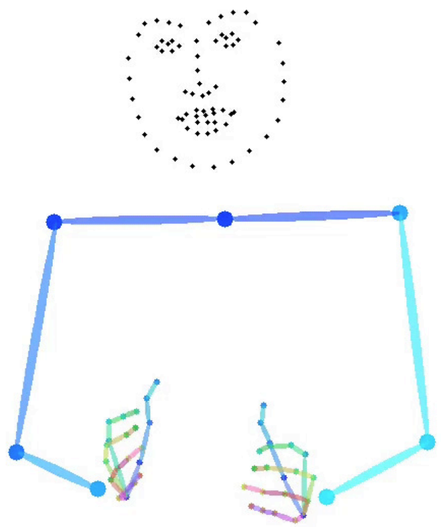
Ground truth

Predicted



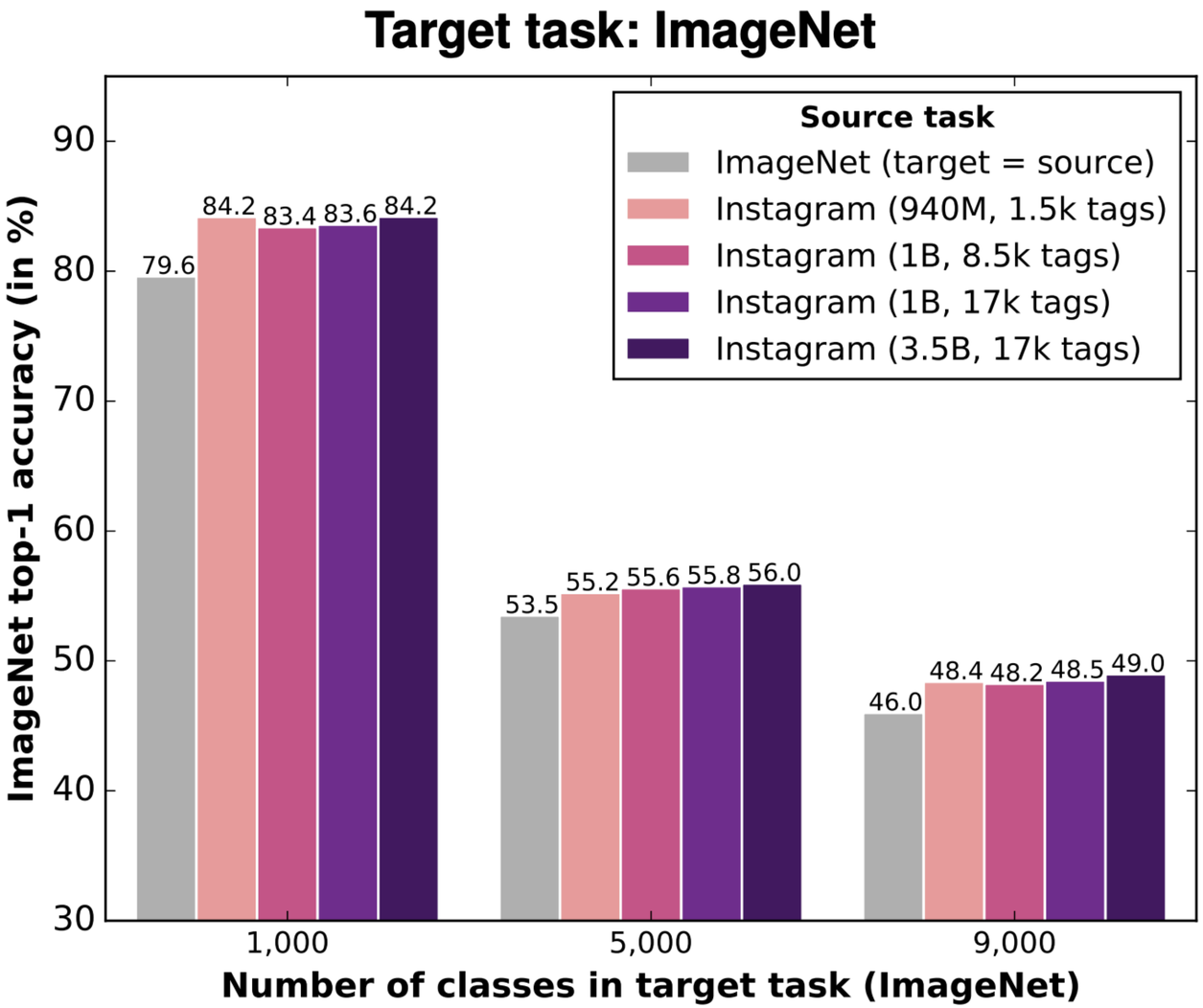
Ground truth

Predicted



Other modalities

Weak supervision from text



Instagram

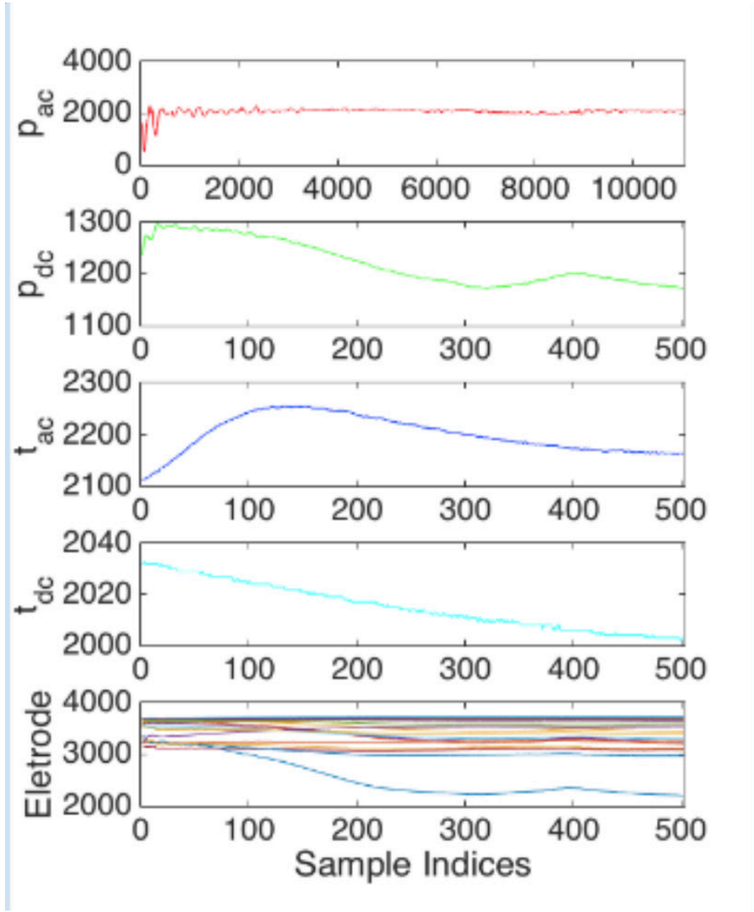
[Mahajan et al. Exploring the Limits of Weakly Supervised Pretraining, 2018]

Post-anesthetization
Performance

From the laboratory of
Dr. Roland Johansson
Dept. of Physiology
University of Umeå, Sweden

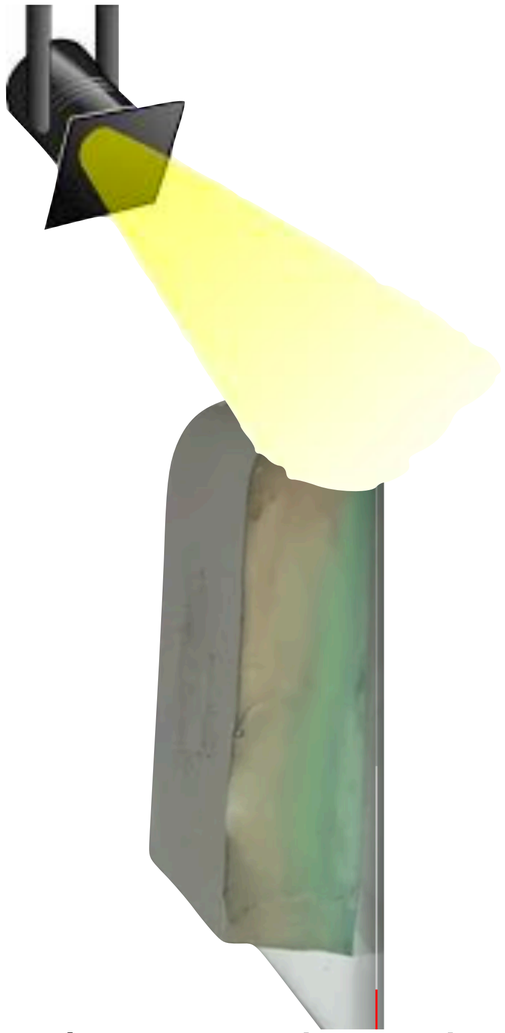
Simple touch sensors

Biotac sensor



Measurements: e.g. force, vibration, temperature

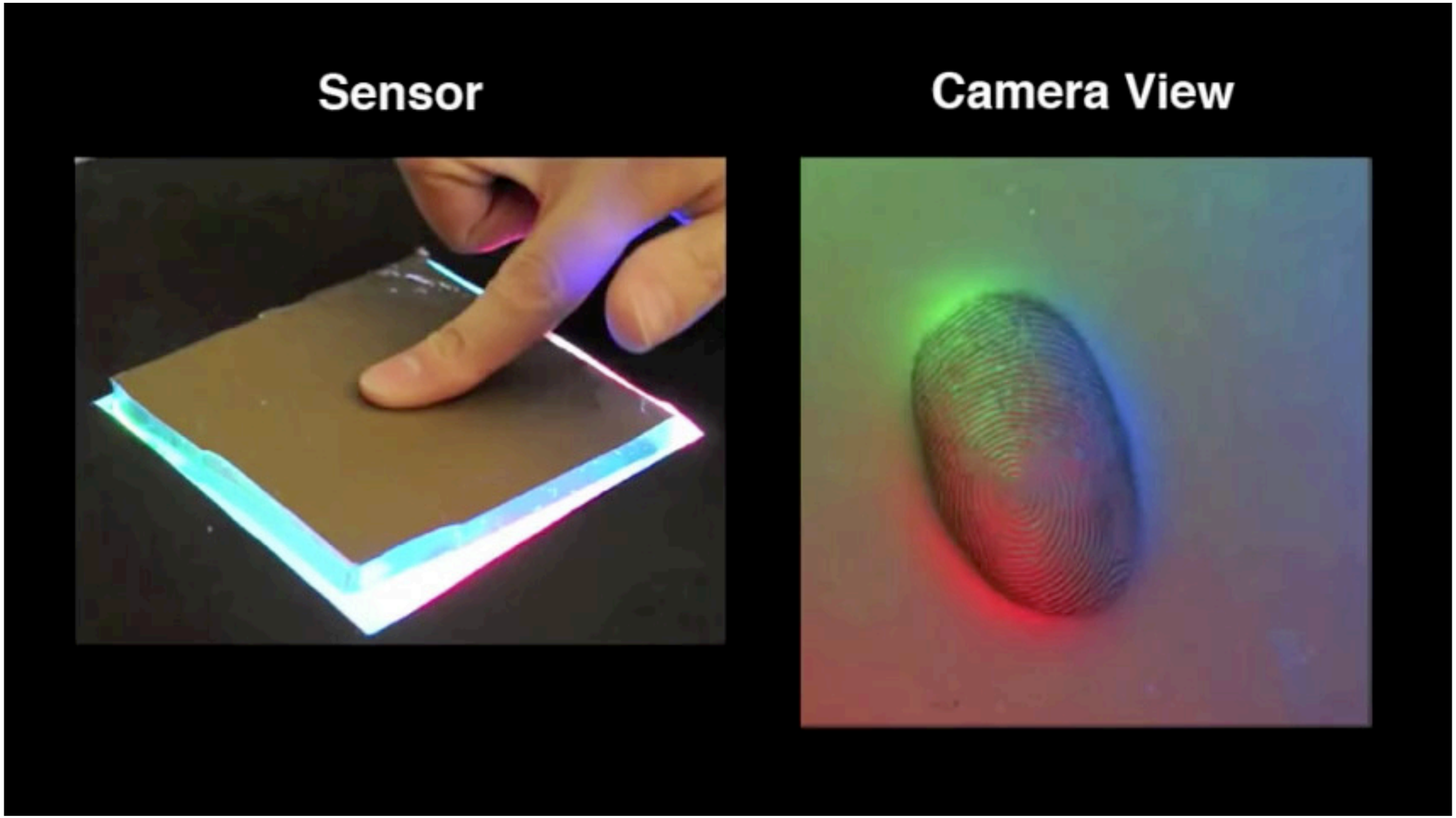
Converting touch to vision



Deform the gel, see the deformation in the camera!

GelSight [Johnson et al. 2009]

Converting touch to vision



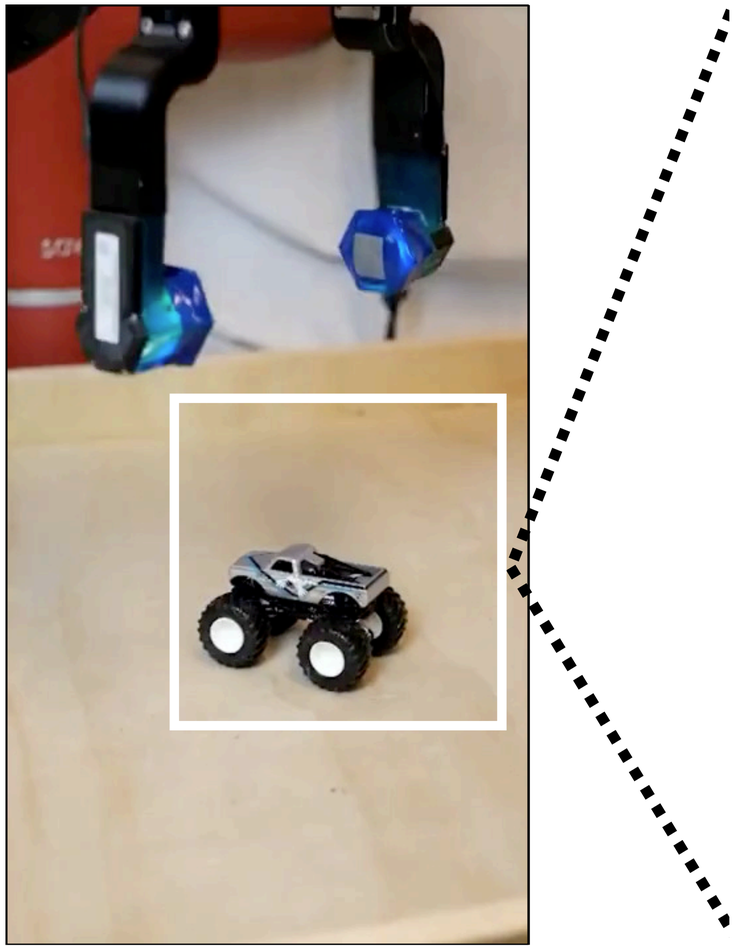
GelSight [Johnson et al. 2009]

Shape-independent Hardness Estimation Using Deep Learning and a GelSight Tactile Sensor

Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens,
Mandayam Srinivasan, Edward Adelson

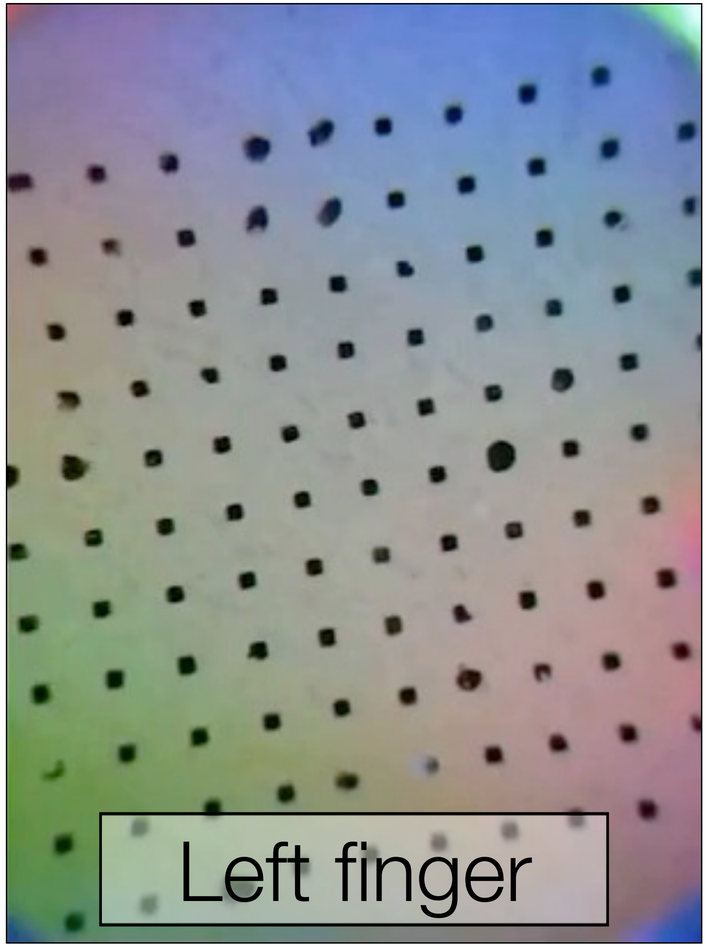
MIT

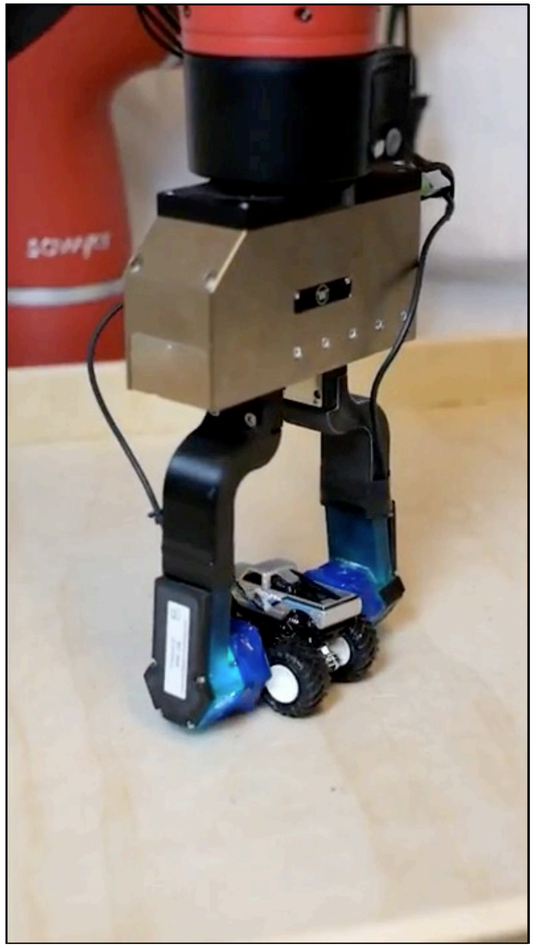
Grasping with vision and touch

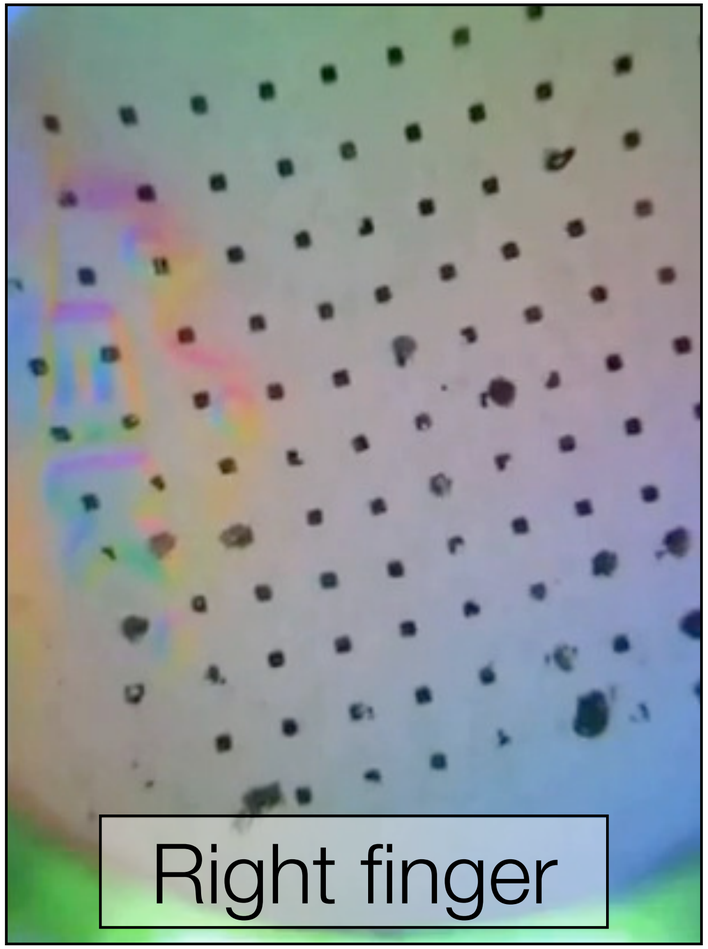


(R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. Adelson, S. Levine. RA-L 2018)
(R. Calandra, A. Owens, M. Upadhyaya, J. Lin, W. Yuan, E. Adelson, S. Levine. CoRL 2018)

Grasping with vision and touch

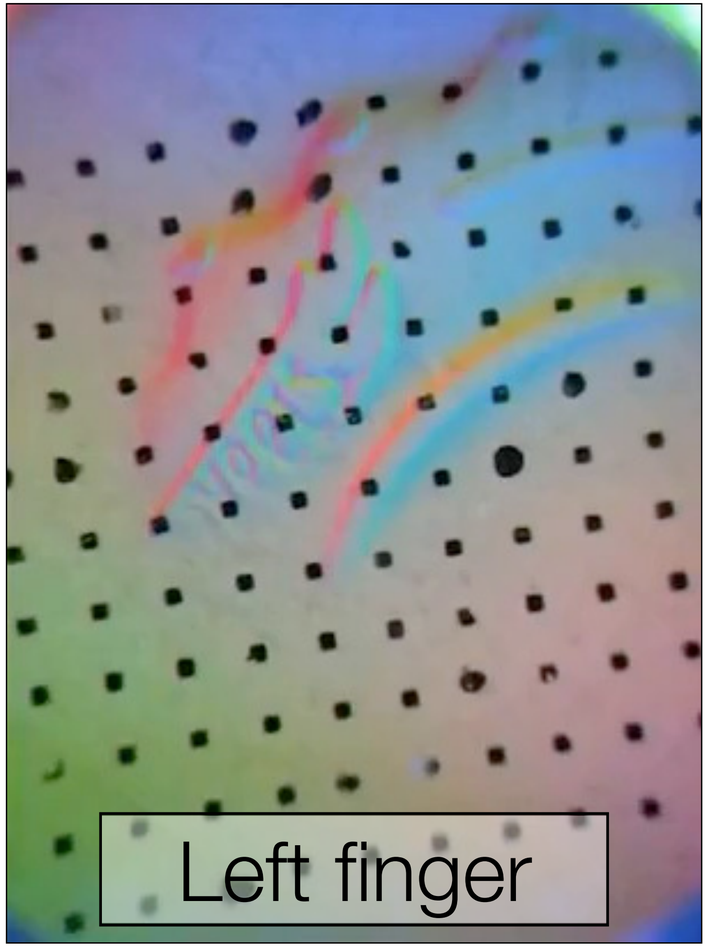


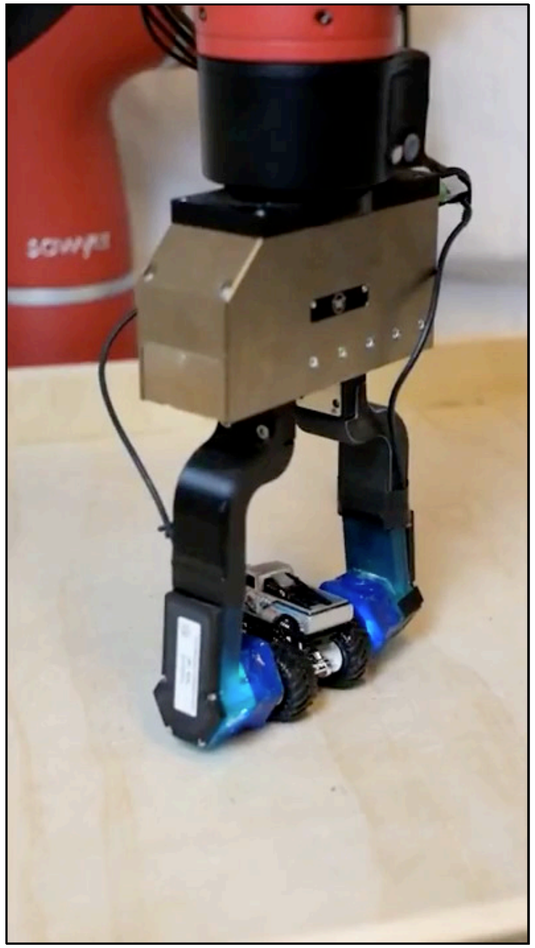


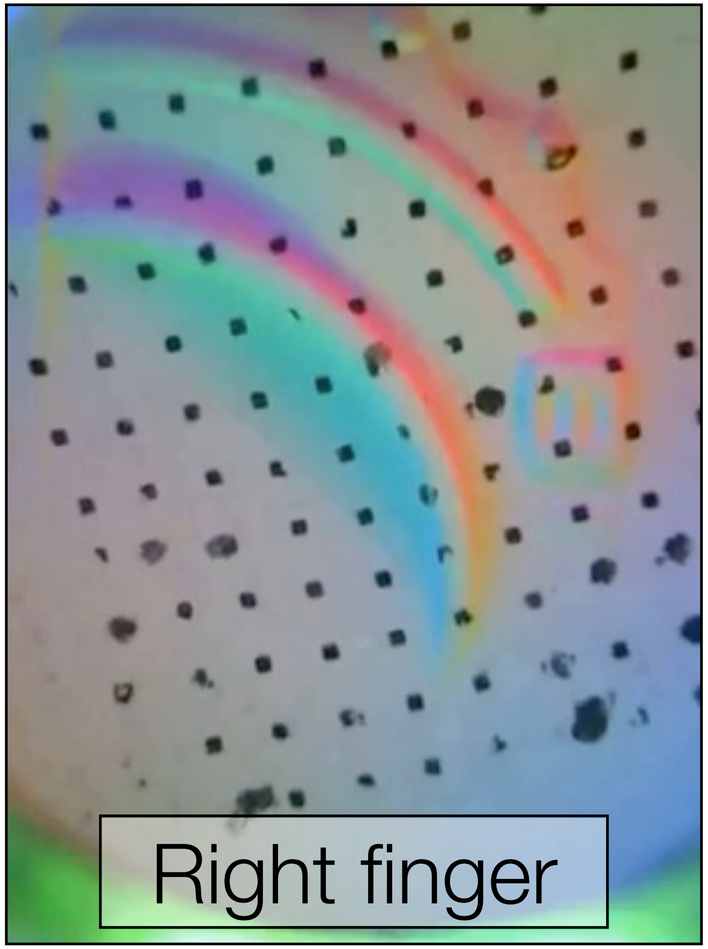


(R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. Adelson, S. Levine. RA-L 2018)
(R. Calandra, A. Owens, M. Upadhyaya, J. Lin, W. Yuan, E. Adelson, S. Levine. CoRL 2018)

Grasping with vision and touch







(R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. Adelson, S. Levine. RA-L 2018)
(R. Calandra, A. Owens, M. Upadhyaya, J. Lin, W. Yuan, E. Adelson, S. Levine. CoRL 2018)

Summary

1. **Multimodal models**: different modalities provide complementary information
2. **Self-supervision**: One sensory signal can provide a learning signal to another
3. **Same methods, different signals:** Can reduce “non-visual” problems into problems that can be solved with neural nets designed for vision