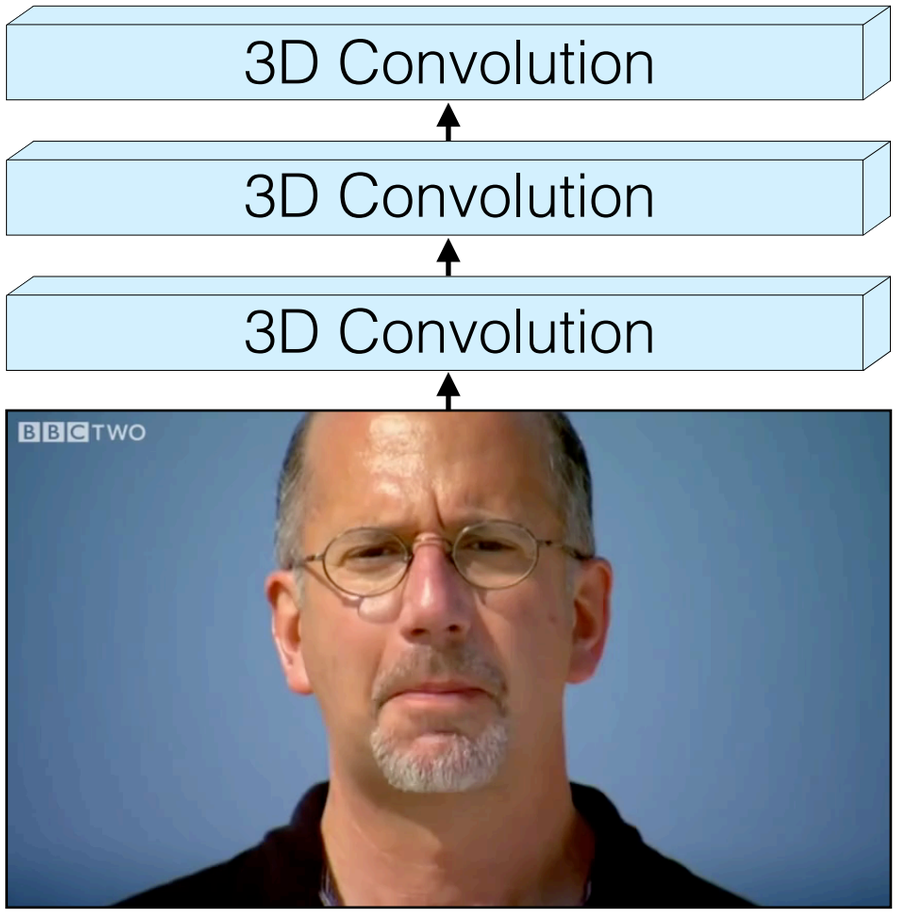
Vision and Sound

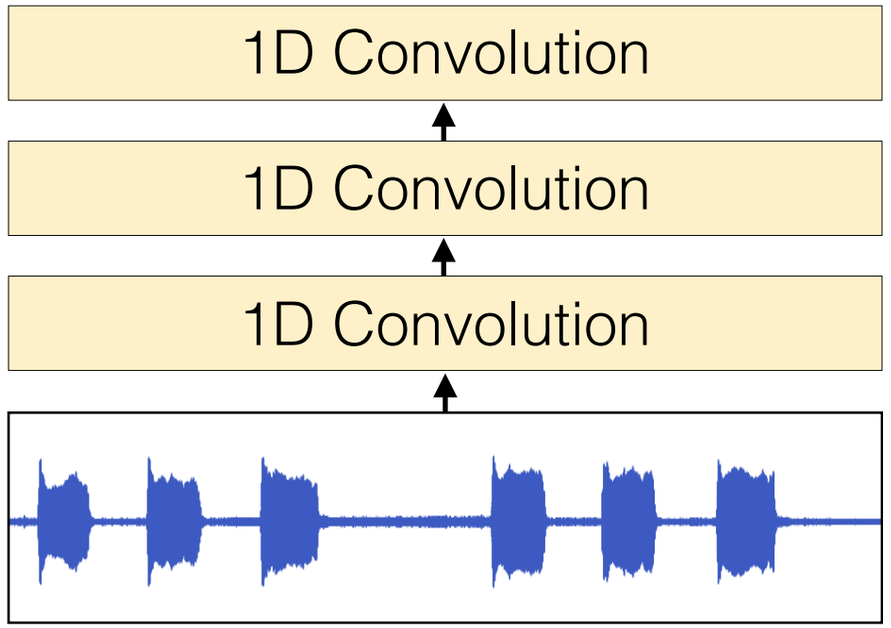
Computer Vision
Fall 2018
Columbia University

Single-modality video representations

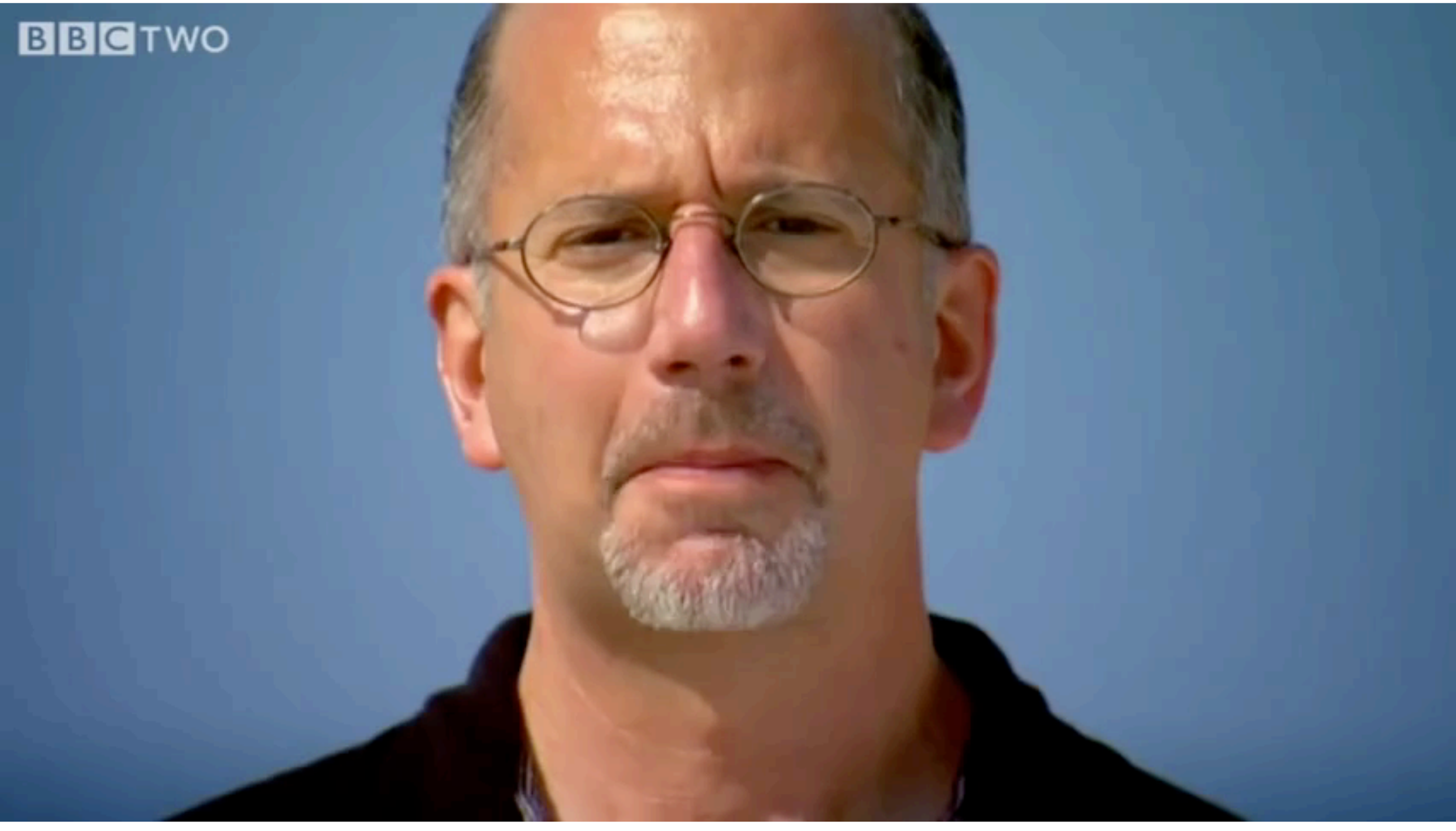
Vision



Hearing



Slide credit: Andrew Owens

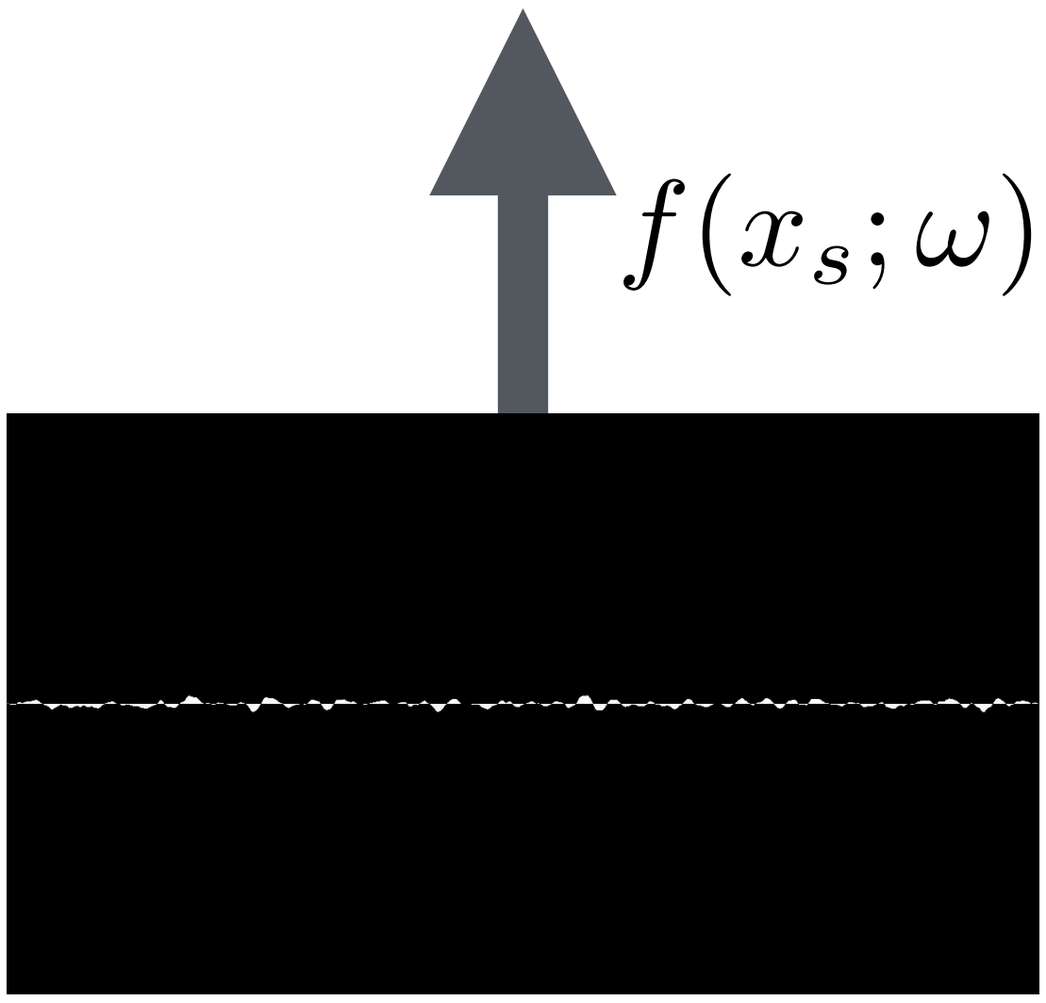


(McGurk 1976)

(McGurk 1976)

Object Recognition

Objects



Sound

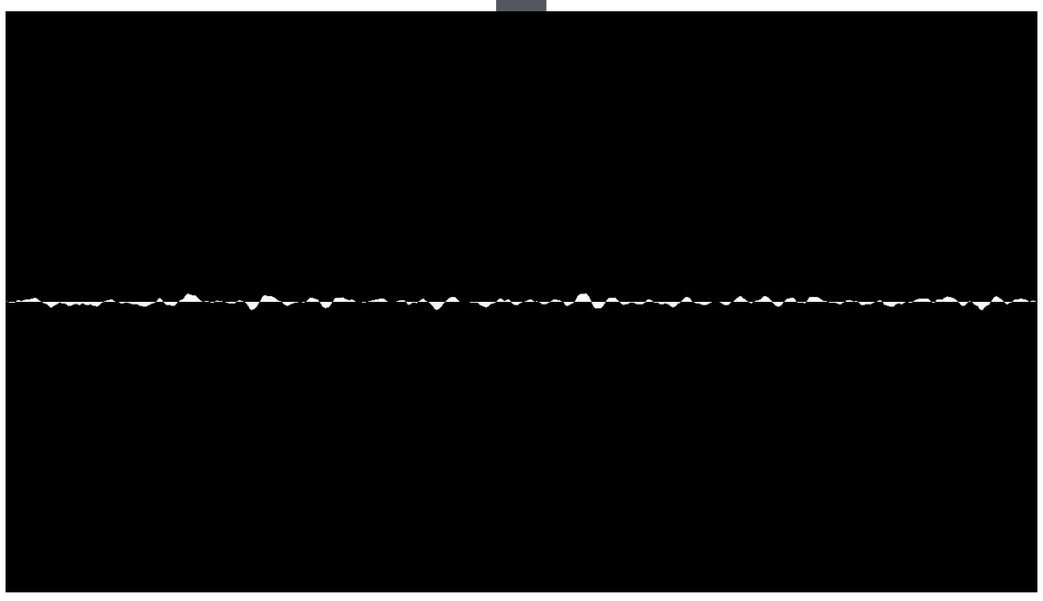
Natural Synchronization

$$\min_f \sum_i D_{\mathrm{KL}}\left(F(x_i) || f(x_i)\right)$$

Lion

$f(x_s;\omega)$

$F(x_v;\Omega)$

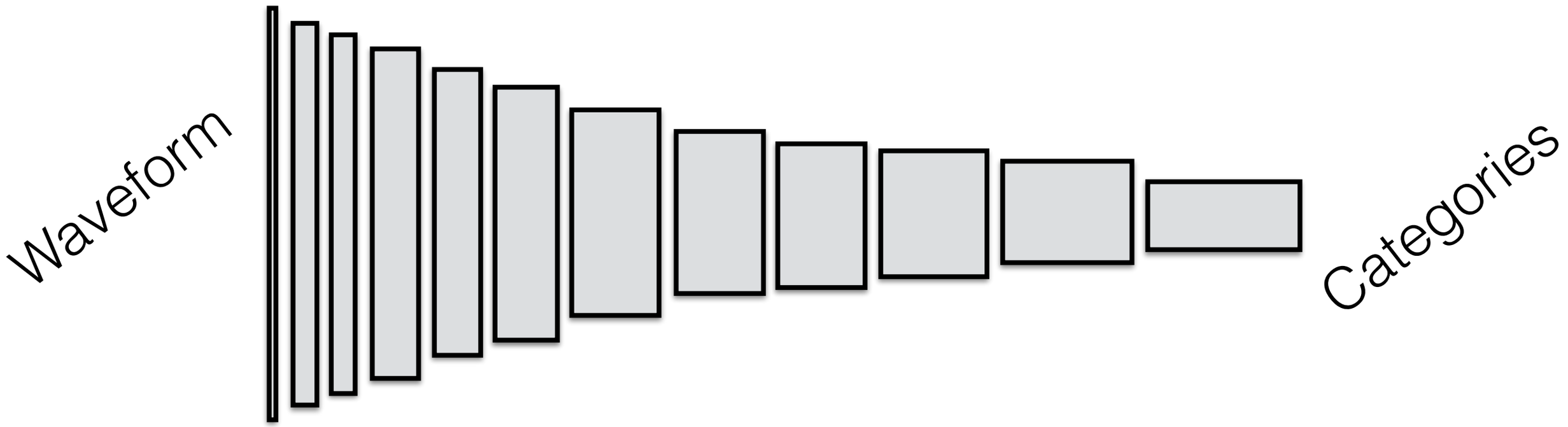


Sound

Vision

Millions of Unlabeled Videos

SoundNet



Convolutional Neural Network

pasture: 18.99%

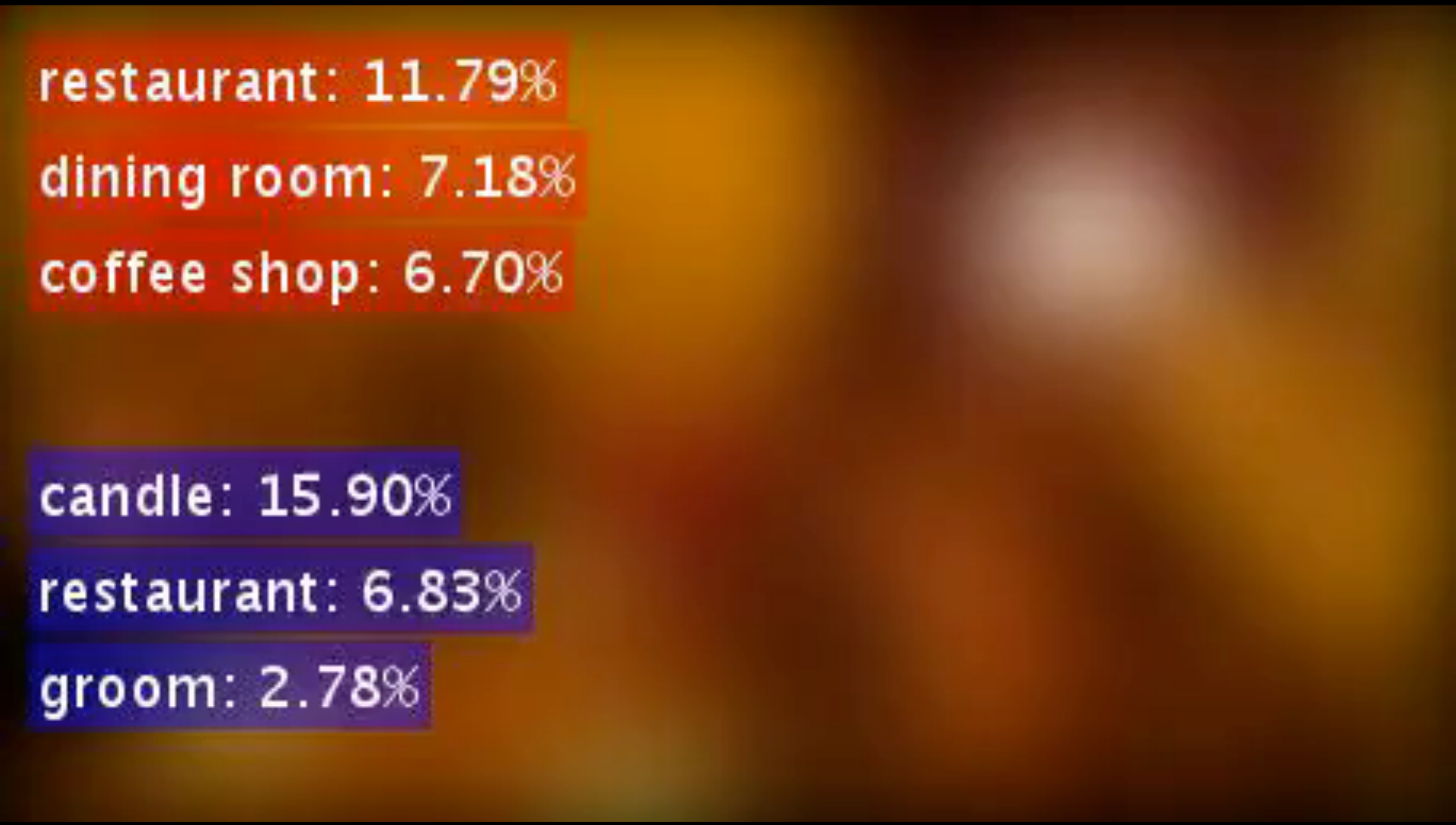
yard: 9.89%

lawn: 7.39%

English springer: 8.65%

Welsh springer spaniel: 2.20%

Border collie: 1.65%


restaurant: 11.79%
dining room: 7.18%
coffee shop: 6.70%
candle: 15.90%
restaurant: 6.83%
groom: 2.78%

Sound Recognition

Classifying sounds in ESC-50

Method	**Accuracy**
Chance	2%
Human Consistency	81%

Sound Recognition

Classifying sounds in ESC-50

Method	**Accuracy**
Chance	2%
SVM-MFCC	39%
Random Forest	44%
CNN, Piczak 2015	64%
Human Consistency	81%

Sound Recognition

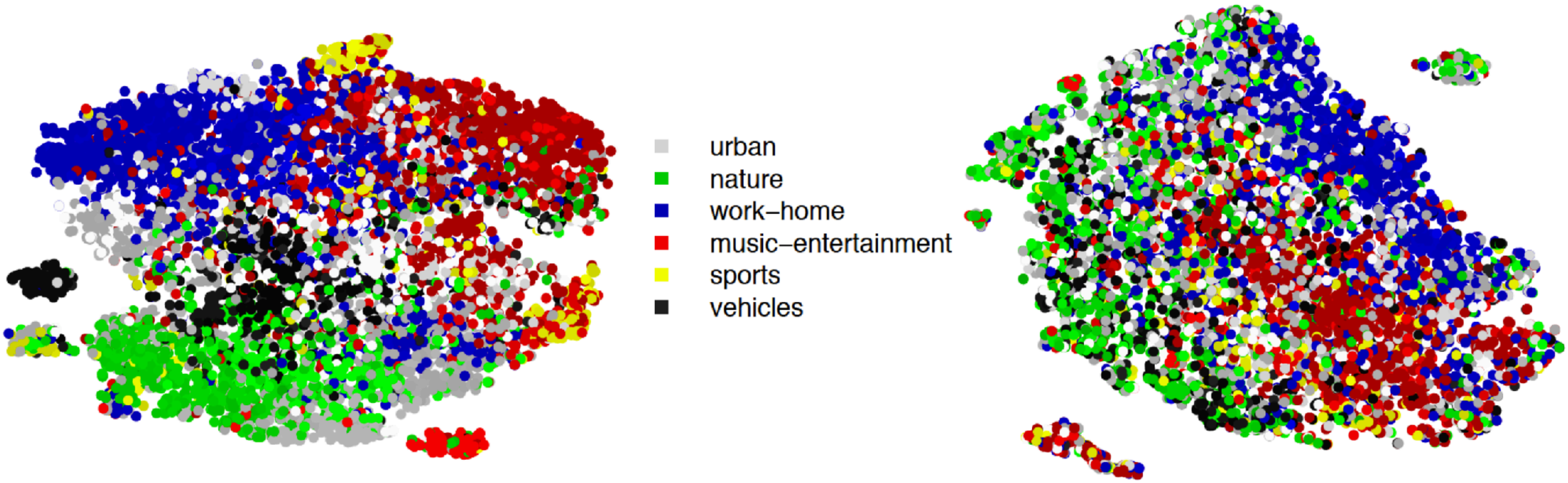
Classifying sounds in ESC-50

Method	**Accuracy**
Chance	2%
SVM-MFCC	39%
Random Forest	44%
CNN, Piczak 2015	64%
SoundNet	**74%**
Human Consistency	81%

10% gain

Vision vs Sound

Low-dimensional embeddings
via Maaten and Hinton, 2007



Vision

Sound

Sensor Power Consumption

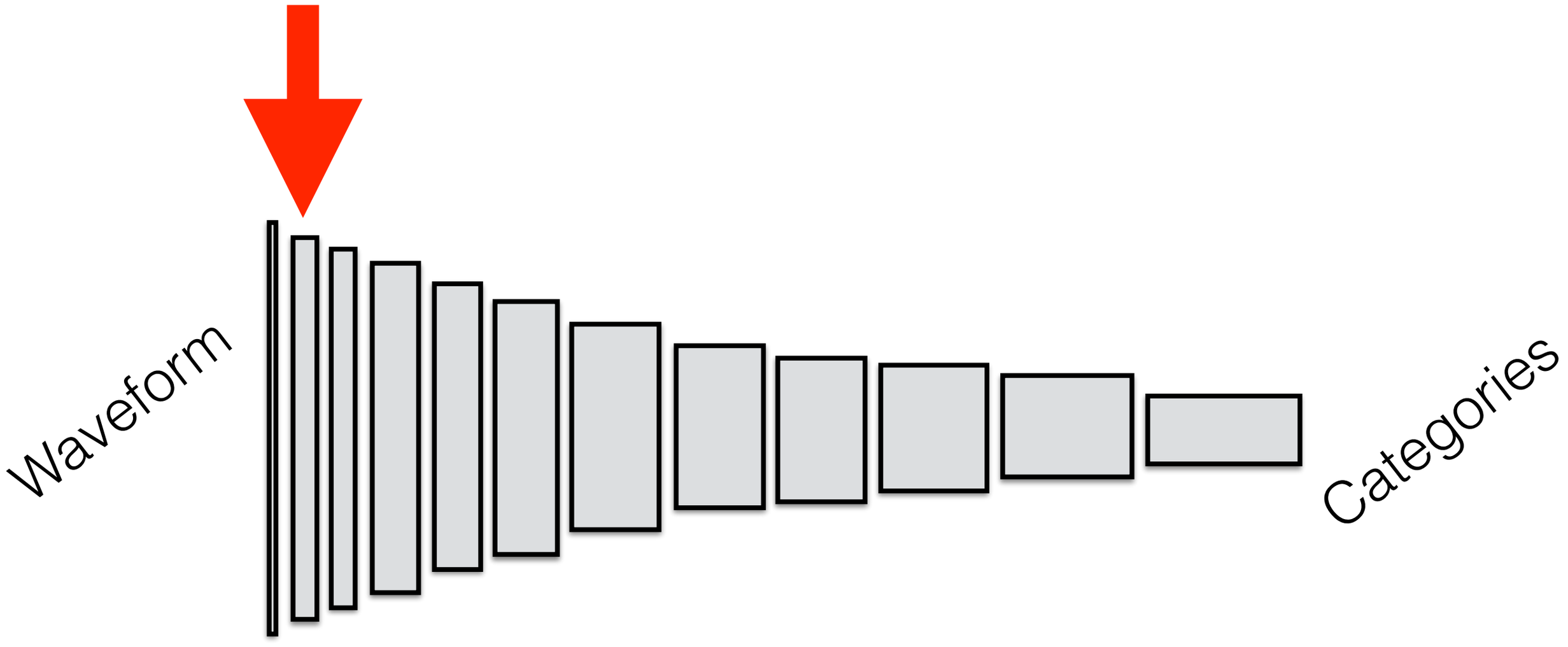
Camera

~1 watt

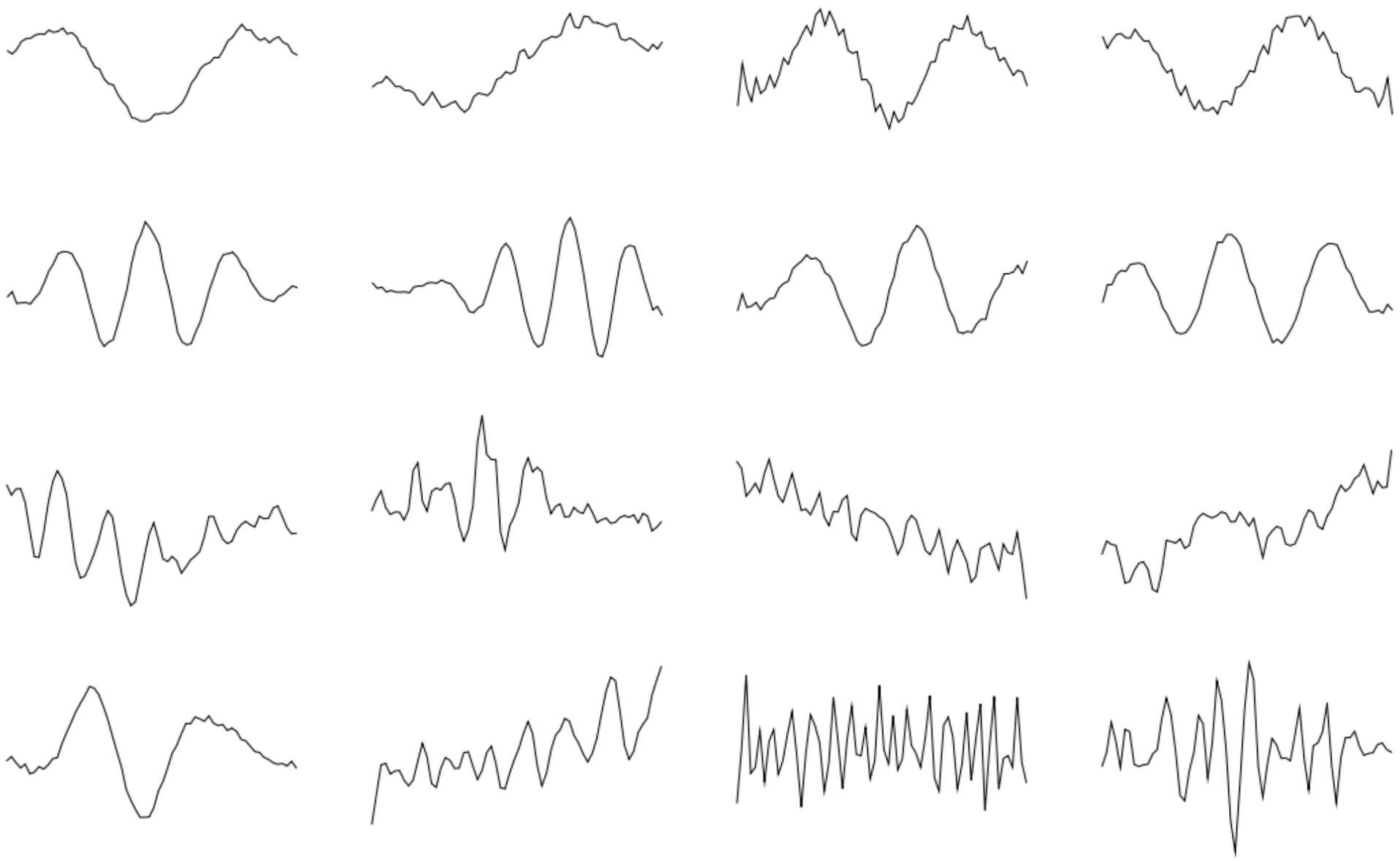
Microphone

~1 milliwatt

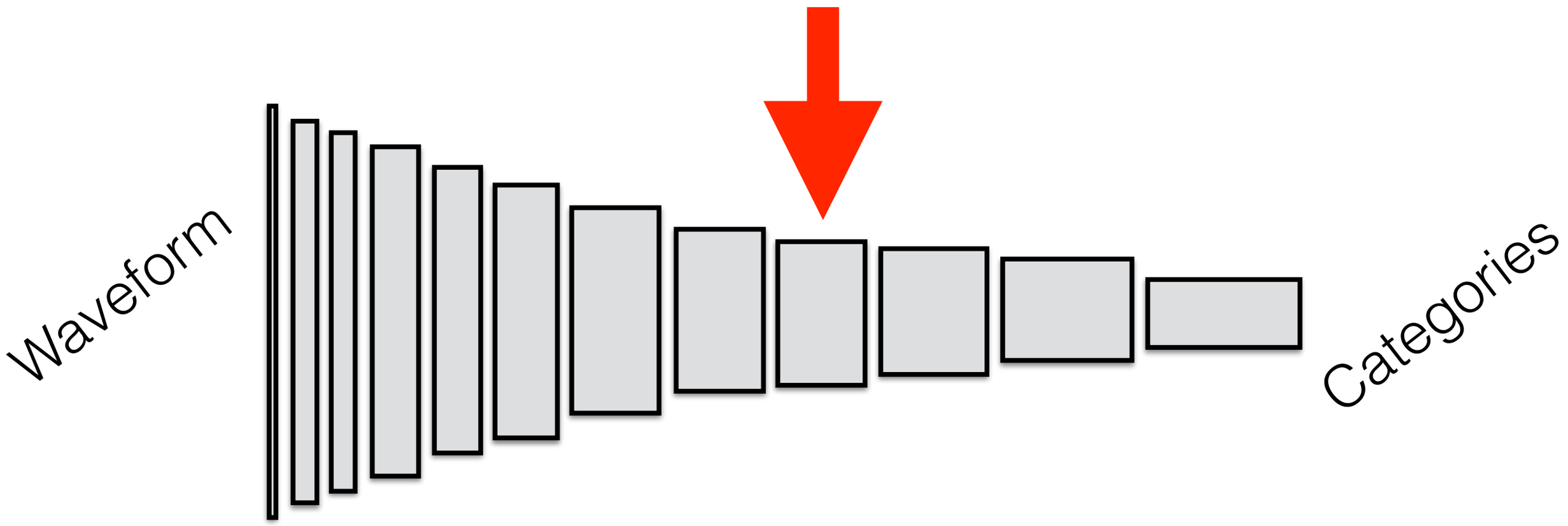
What does it learn?



Layer 1



What does it learn?



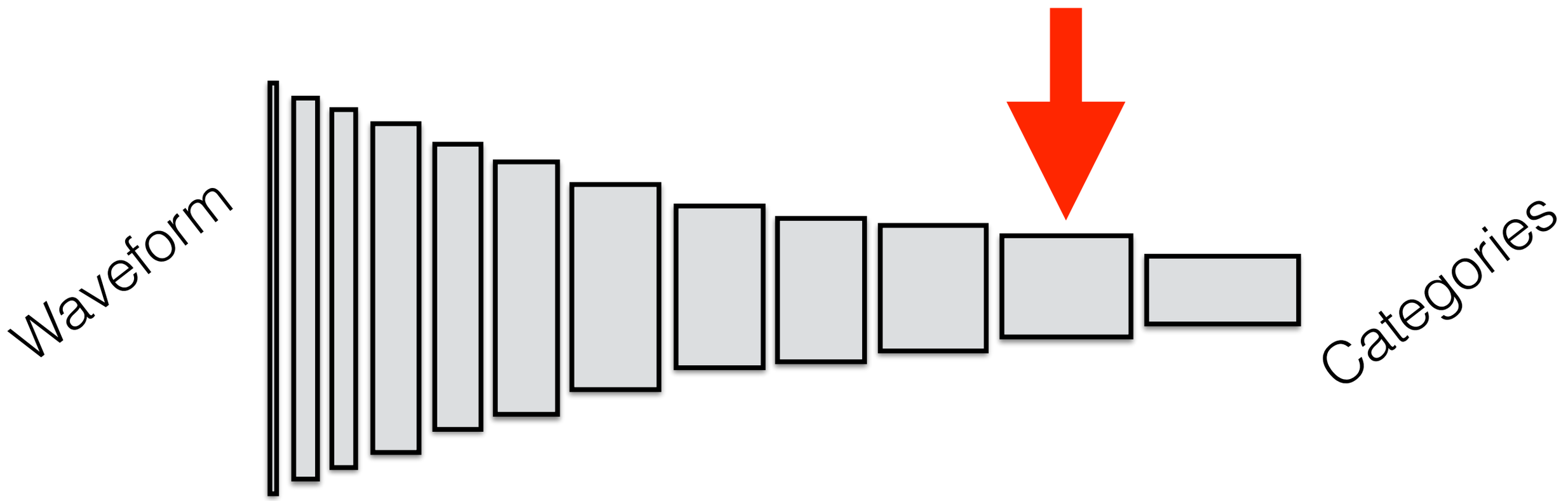
Layer 5

Smacking-like

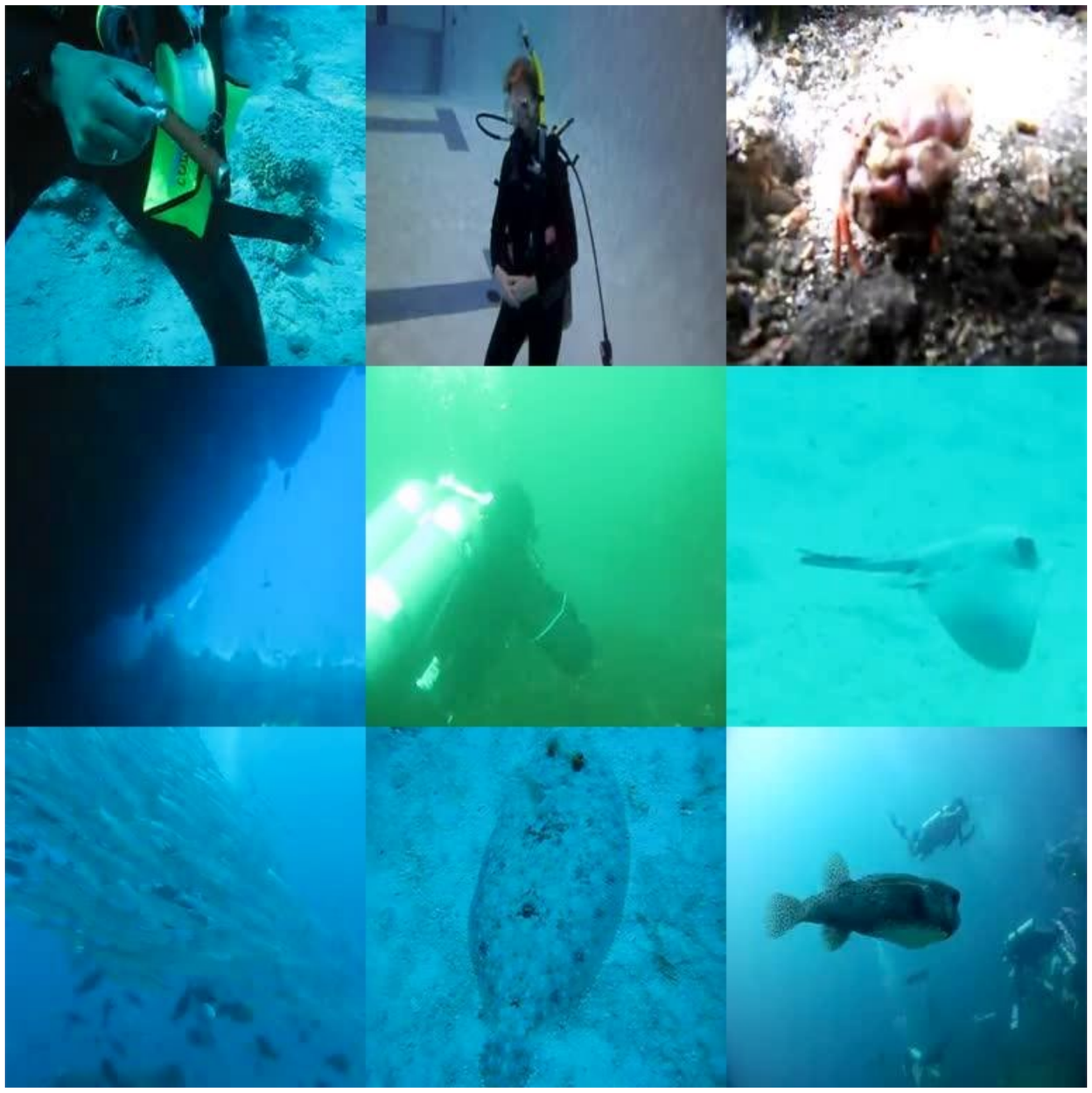
Layer 5

Chime-like

What does it learn?



Layer 7



Scuba-like

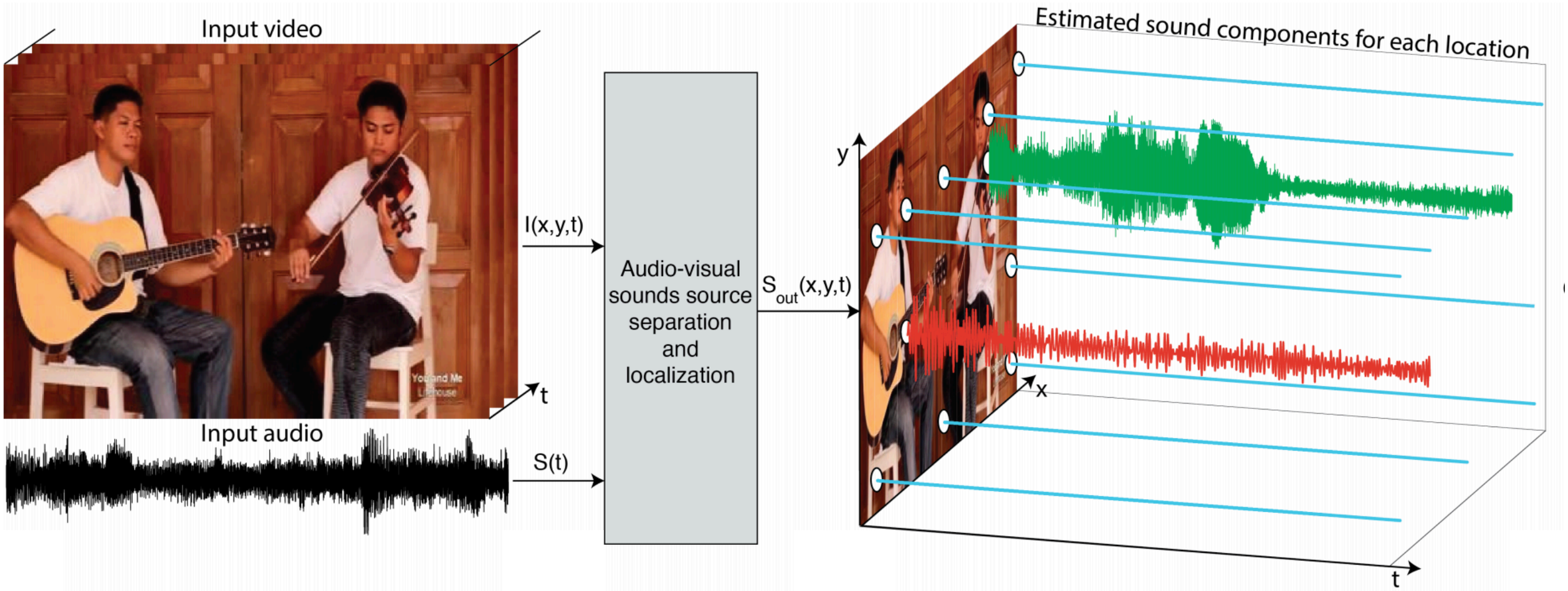
Layer 7

Parents-like

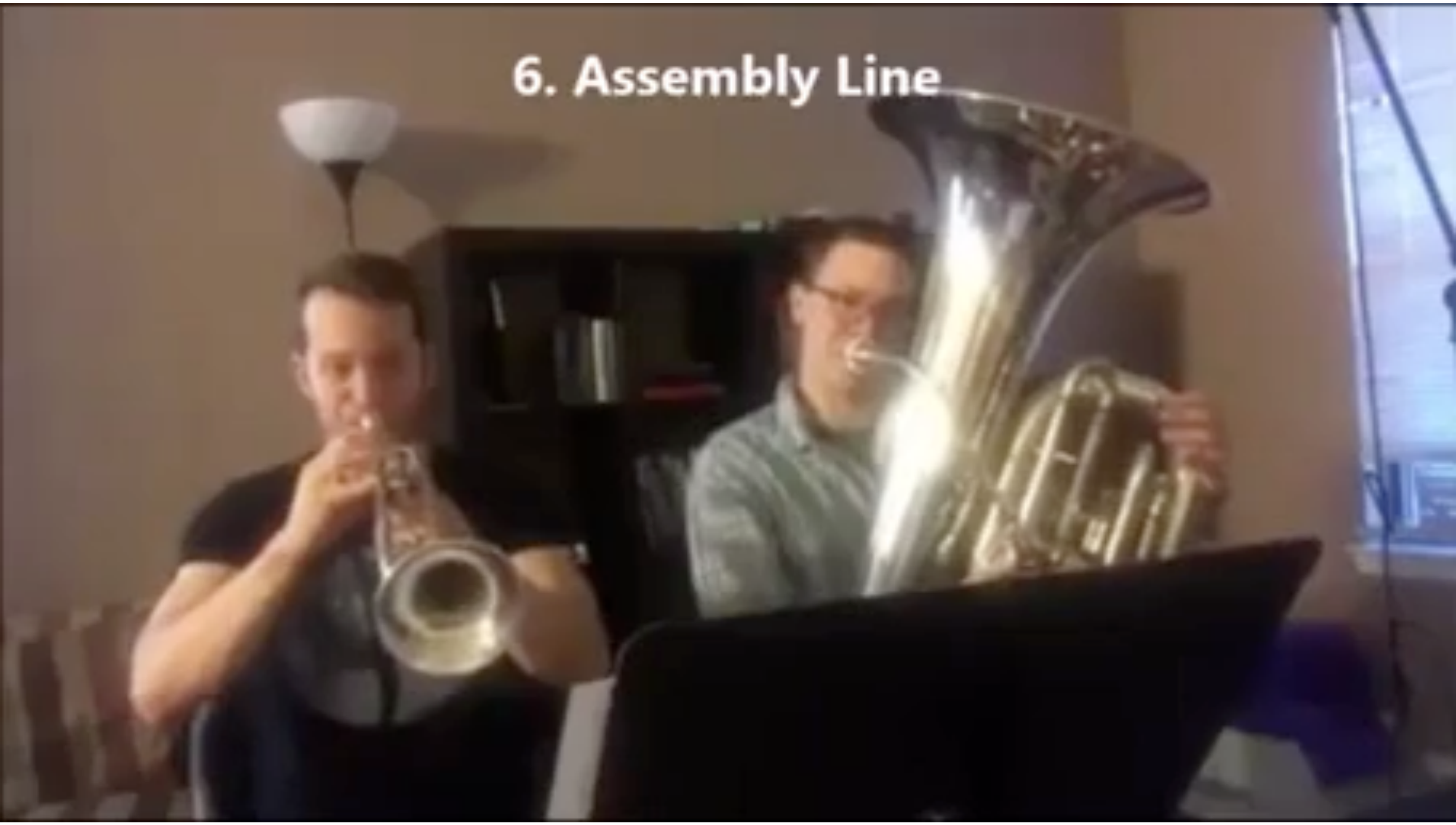
Audiovisual Grounding

Which regions are making which sounds?

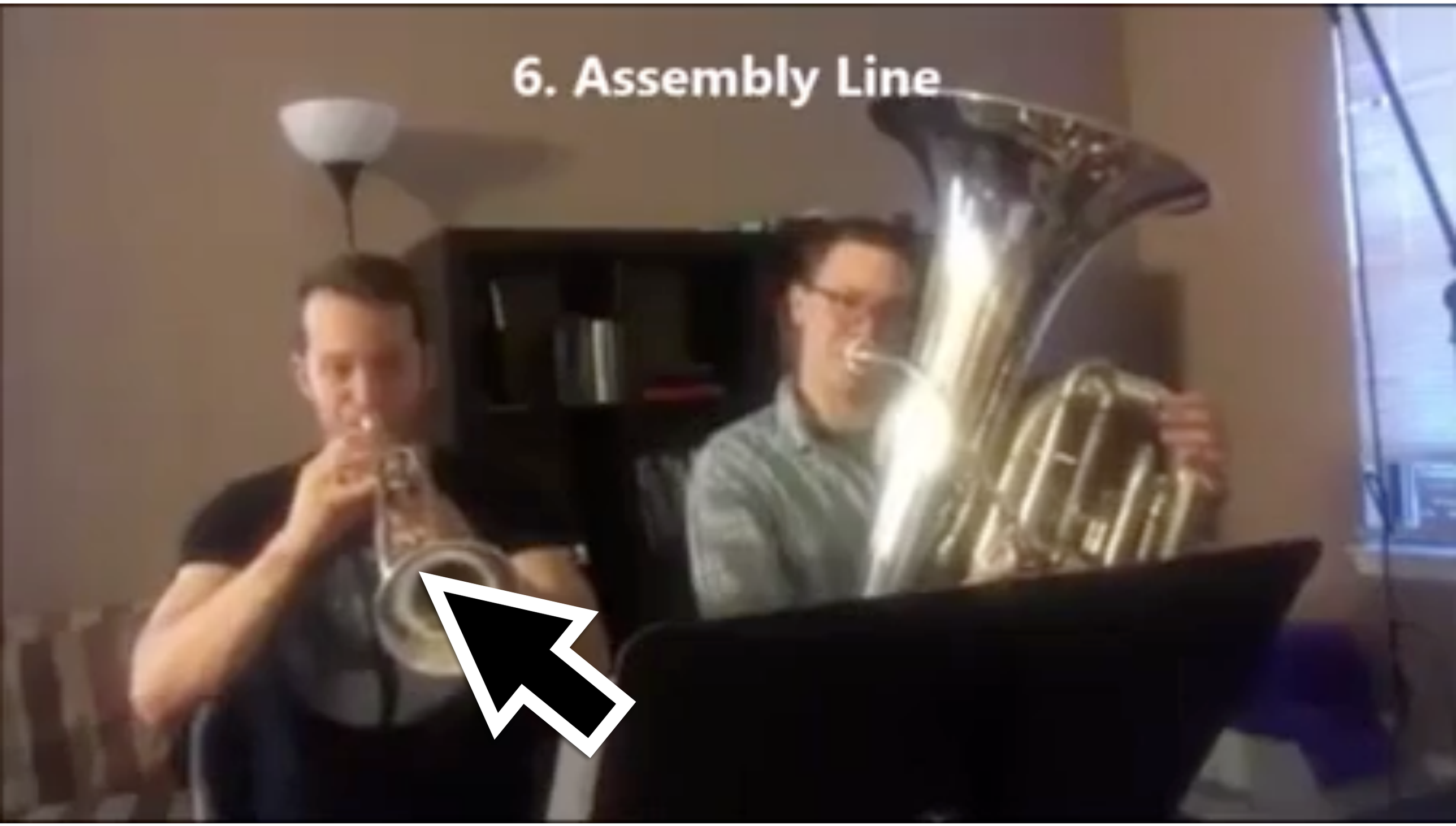
Audiovisual Grounding



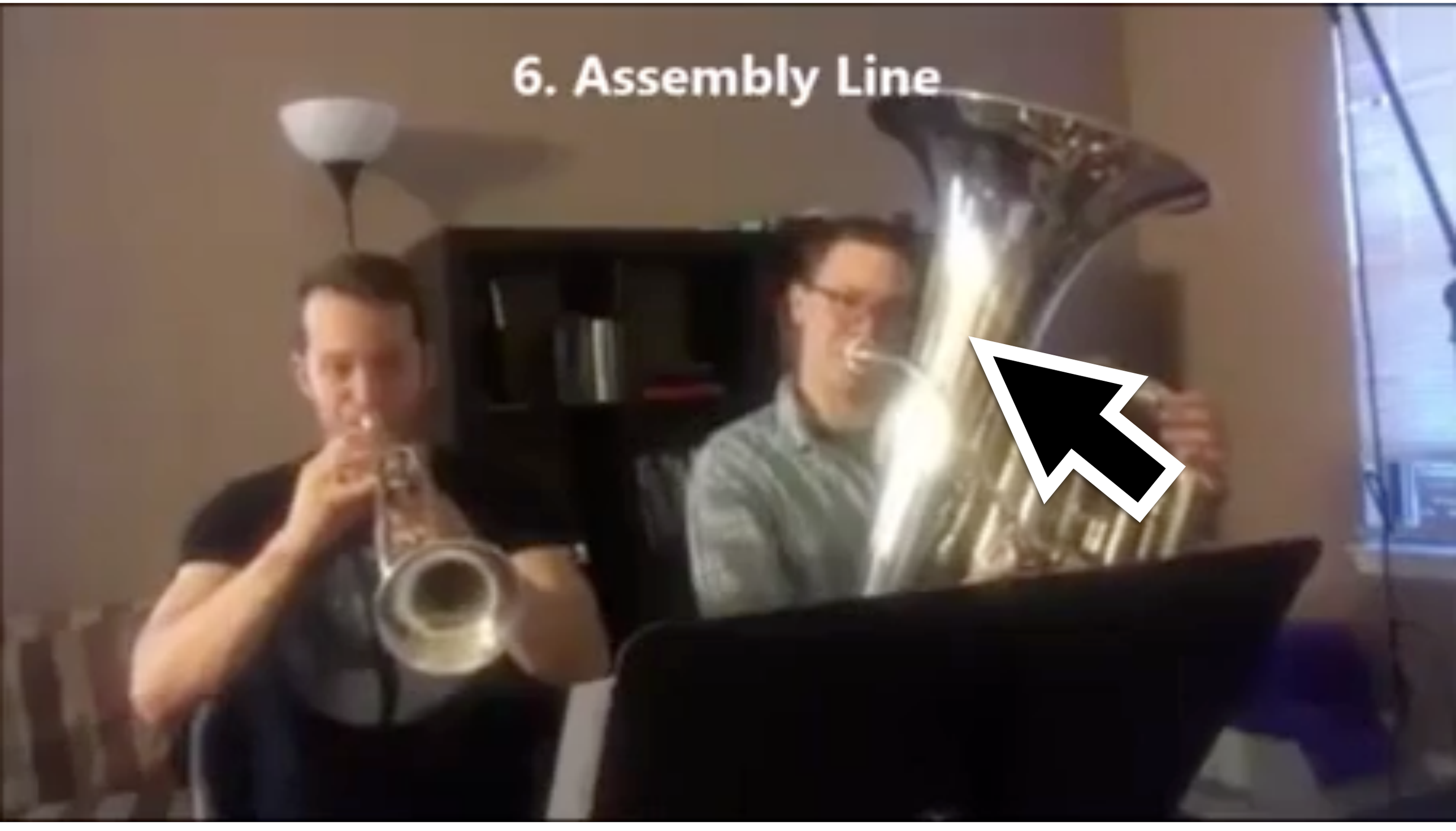
Which objects make which sounds?



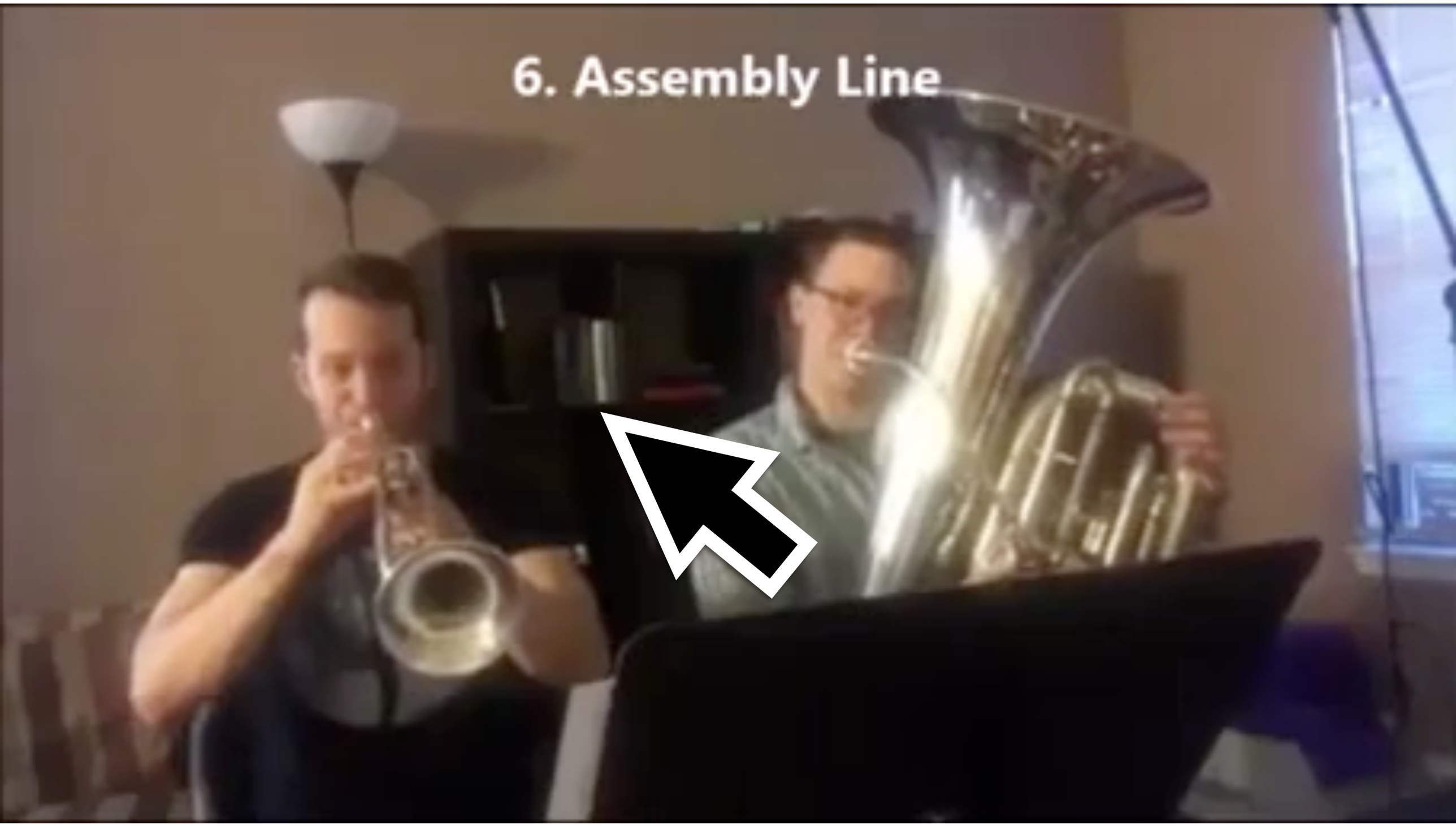
The sound of clicked object



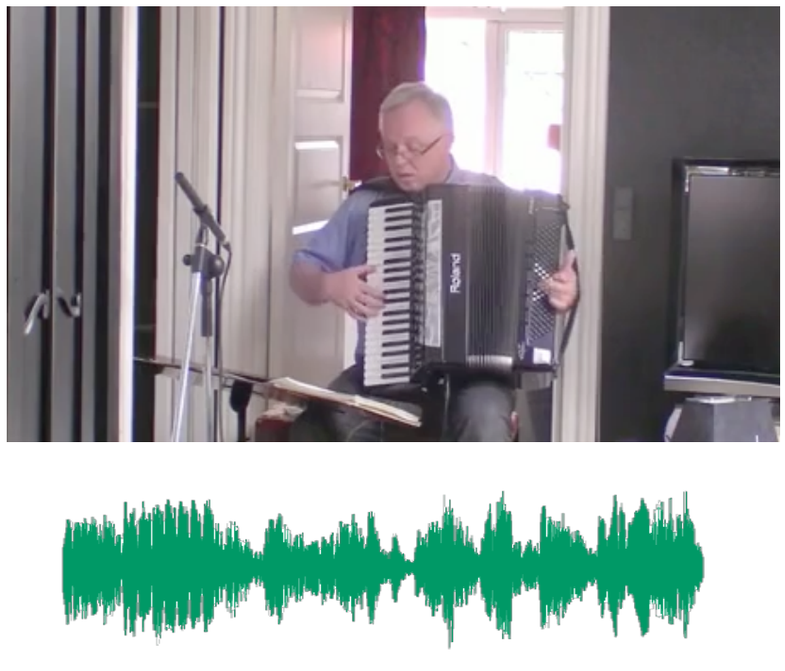
The sound of clicked object

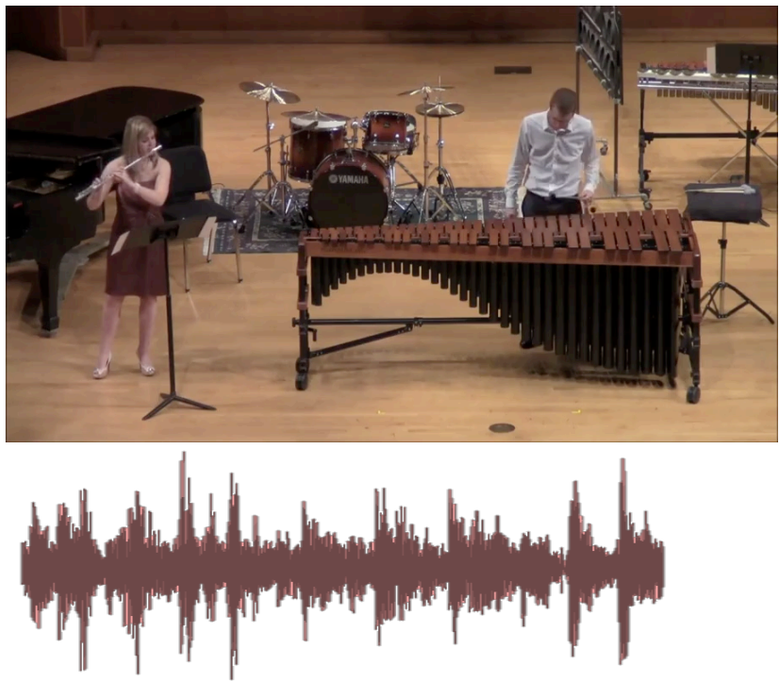


The sound of clicked object

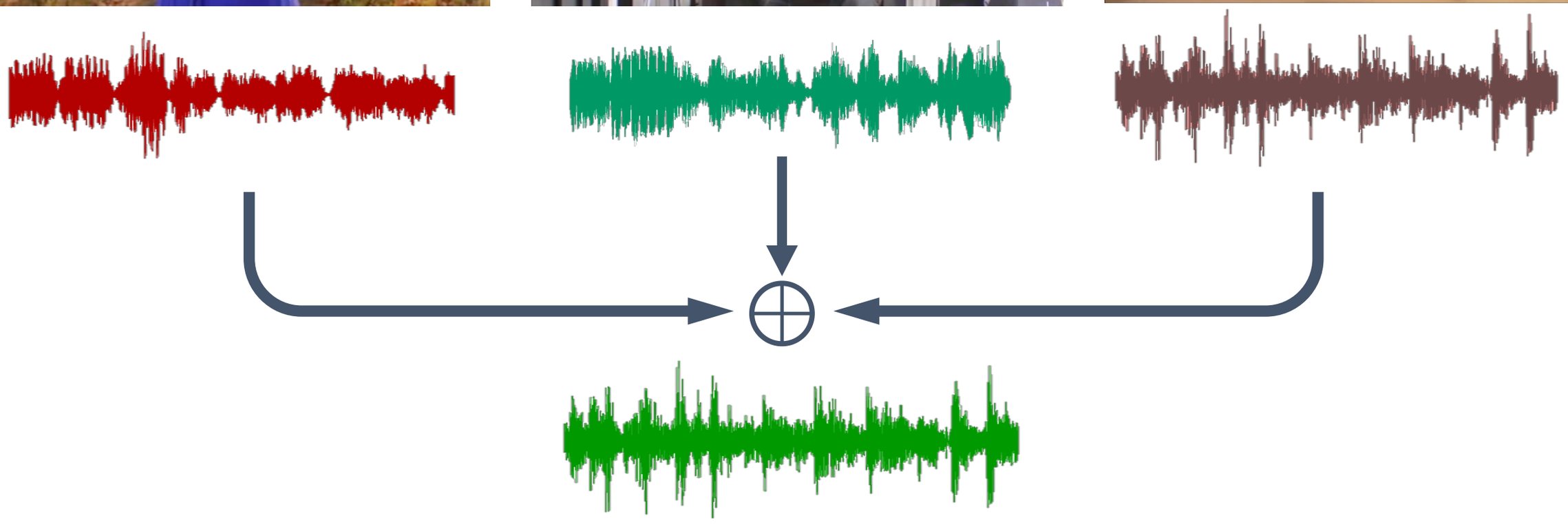


Collect unlabeled videos





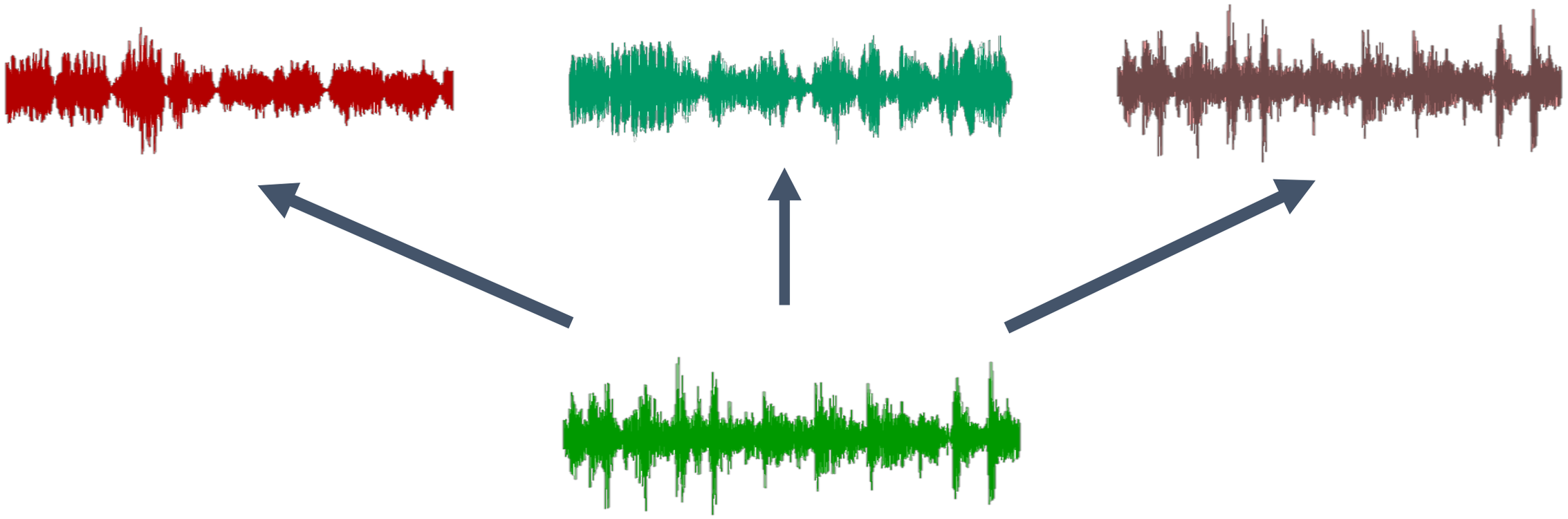
Mix Sound Tracks



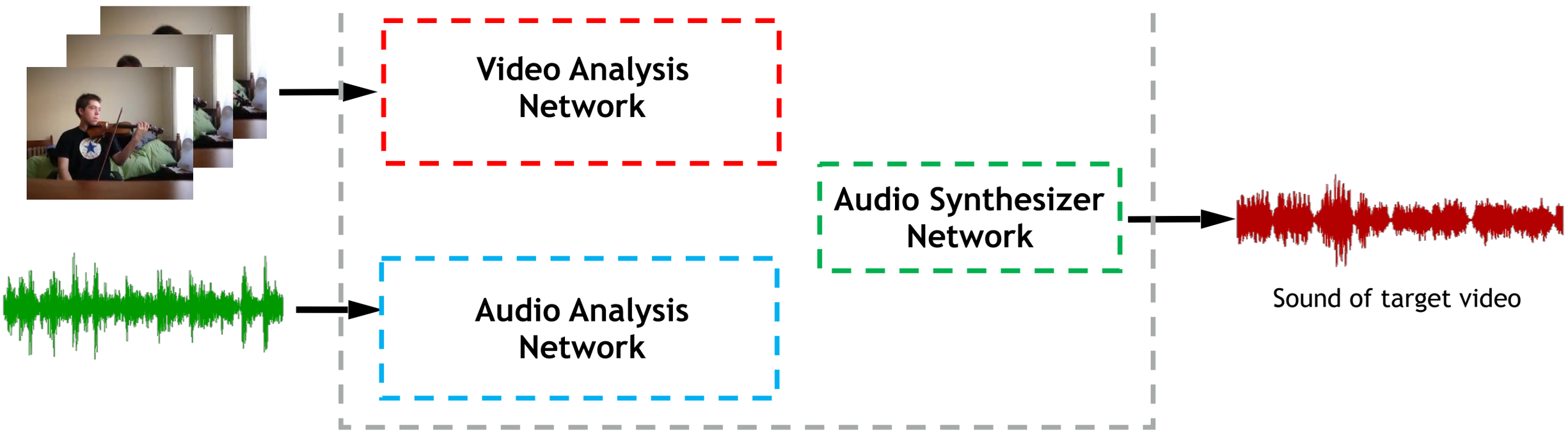
How to recover originals?

Audio-only:

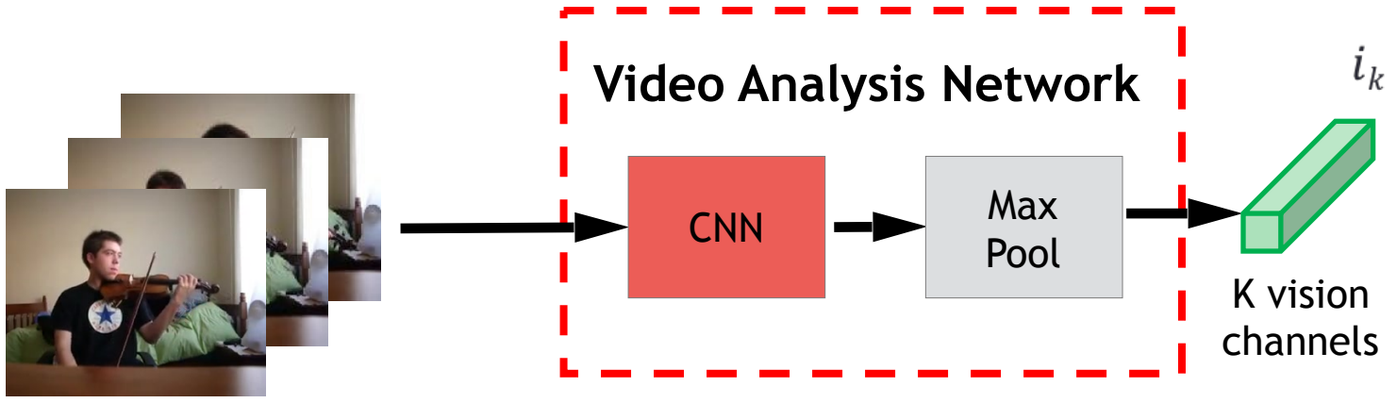
- Ill-posed
- permutation problem



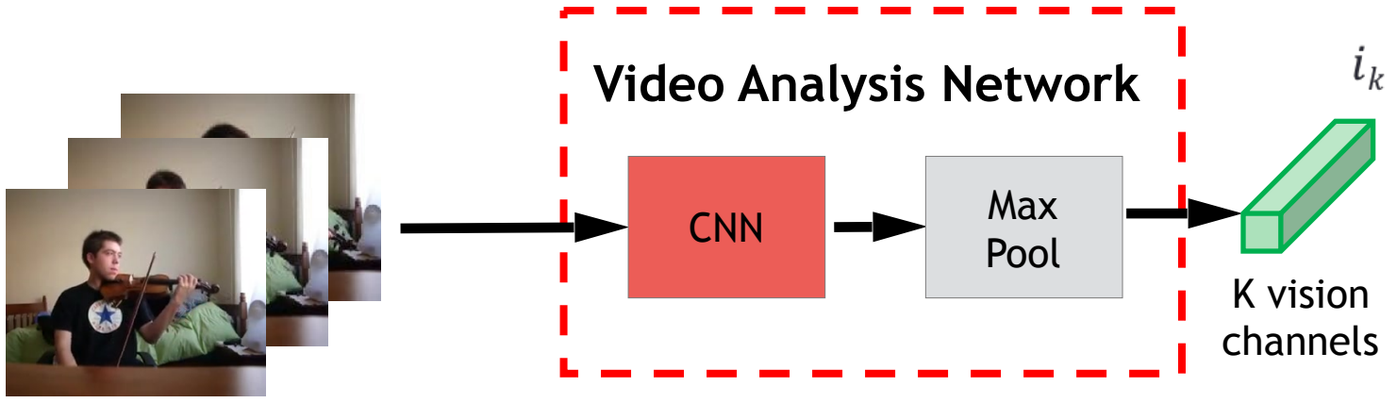
Vision can help

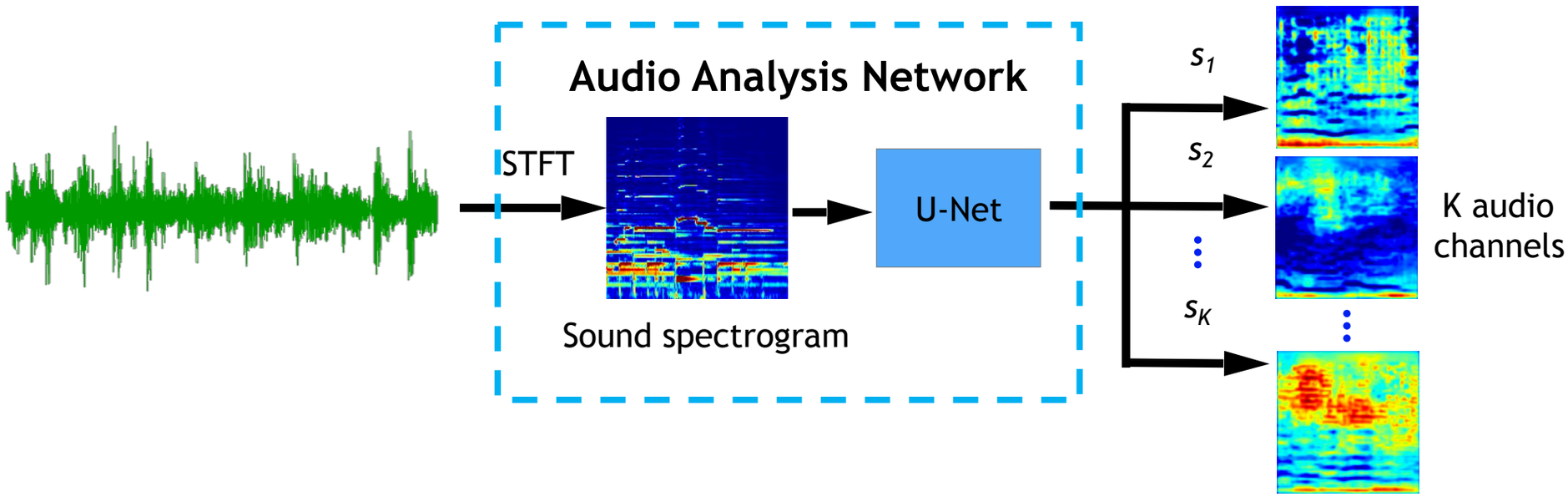


Audiovisual Model

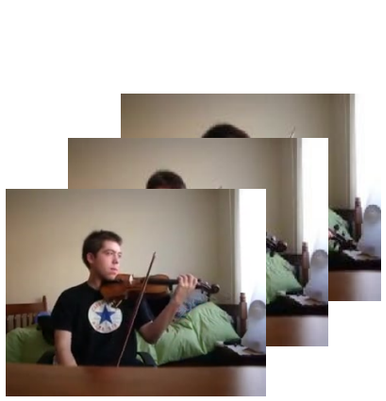


Audiovisual Model





Audiovisual Model



Video Analysis Network

CNN

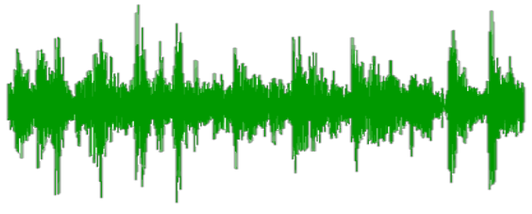
Max Pool

i_k

K vision channels

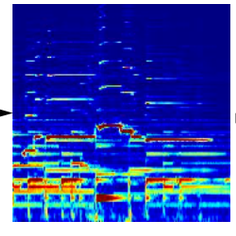
Audio Synthesizer Network

Sound of target video



Audio Analysis Network

STFT



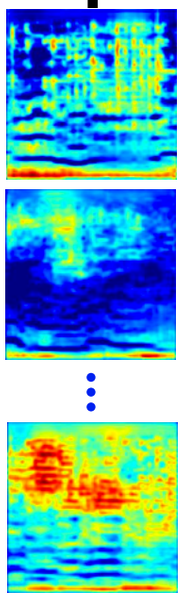
Sound spectrogram

U-Net

s_1

s_2

s_K

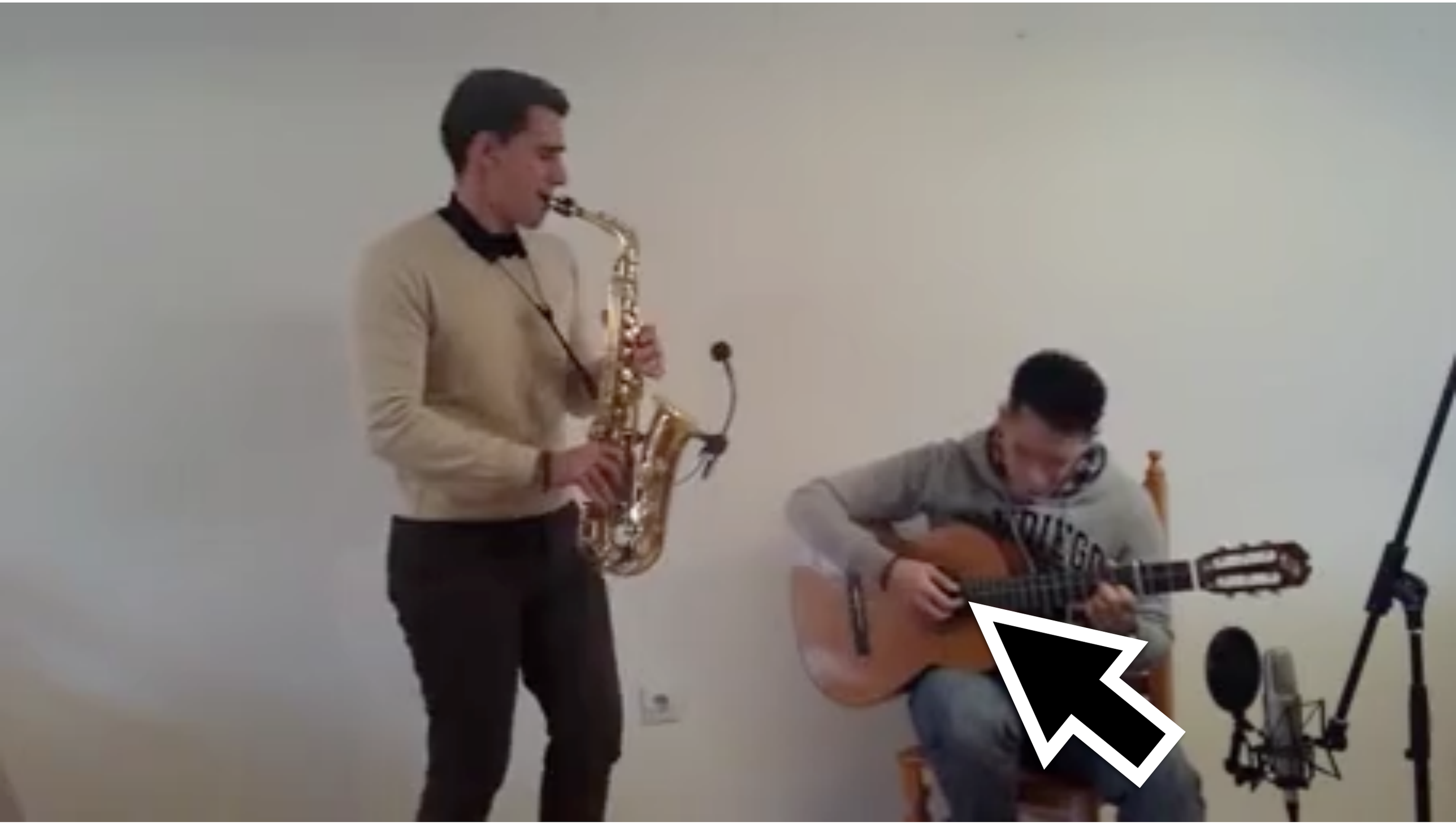


K audio channels

Original Audio

What does this sound like?

What does this sound like?



What does this sound like?

What regions are making sound?

Original Video

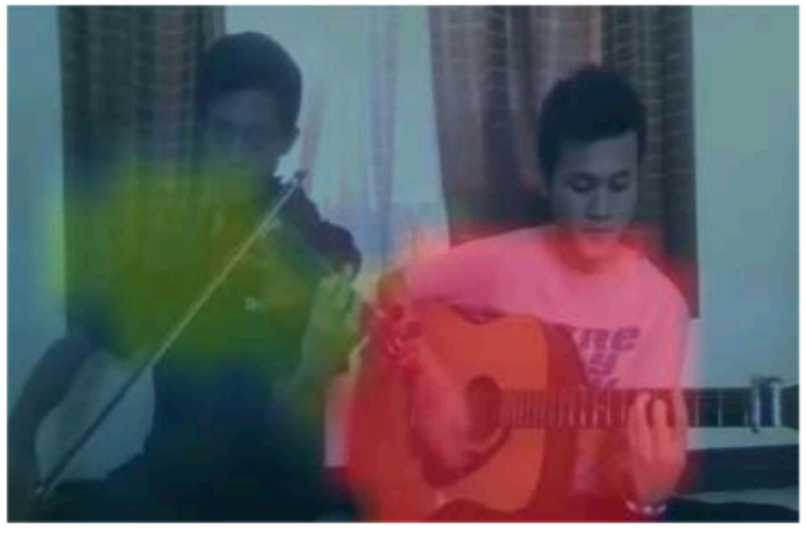
Estimated Volume

What sounds are they making?

Original Video








Embedding (projected and visualized as color)

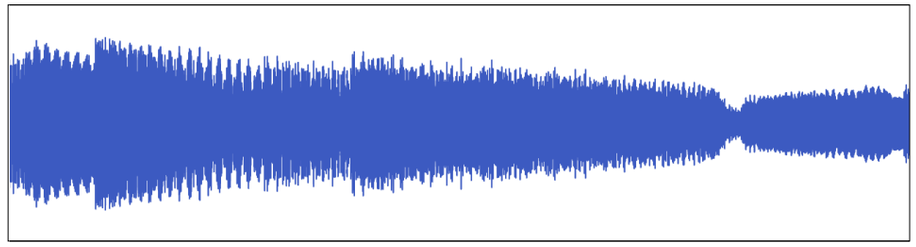
Adjusting Volume

Volume 1

Play

Volume 2

Learning audio-visual correspondences

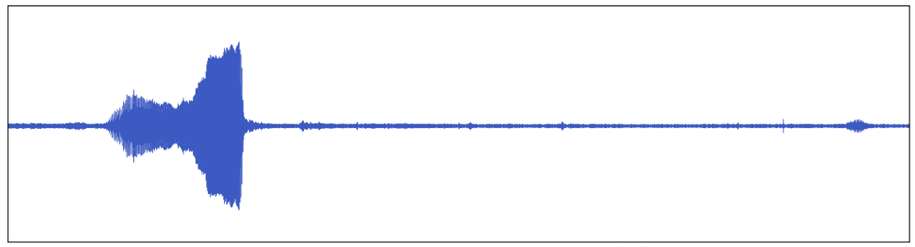


, → **real** or fake?

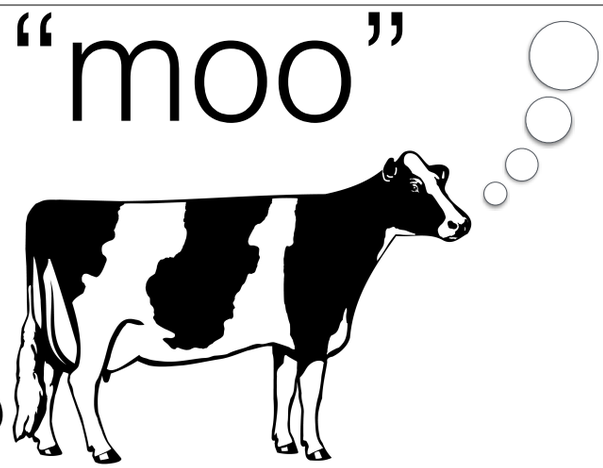
Slide credit: Andrew Owens

Learning audio-visual correspondences

,



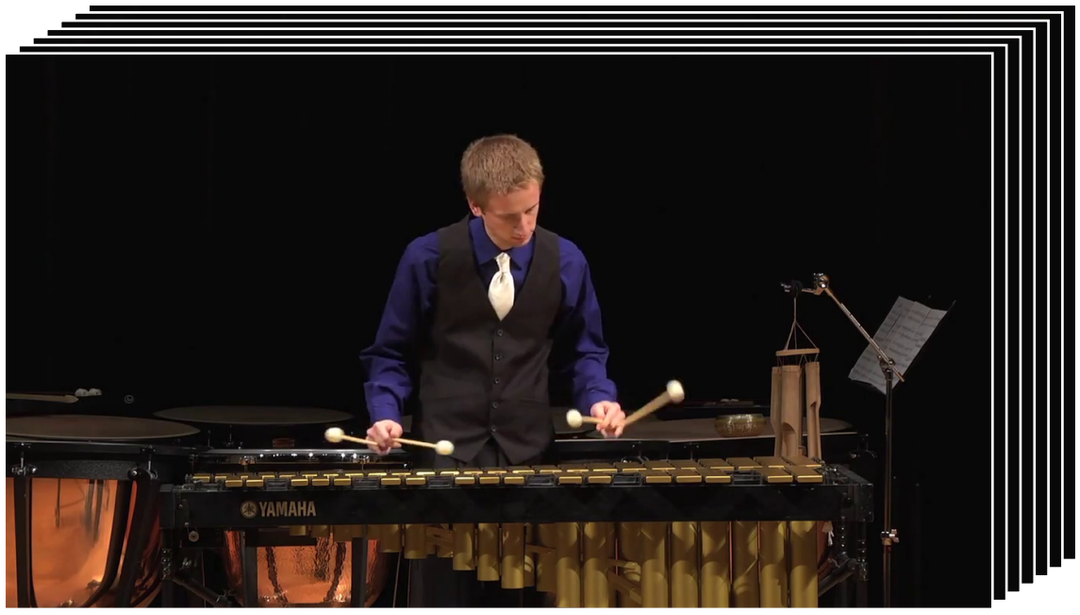
“moo”



→ real or **fake**?

Slide credit: Andrew Owens

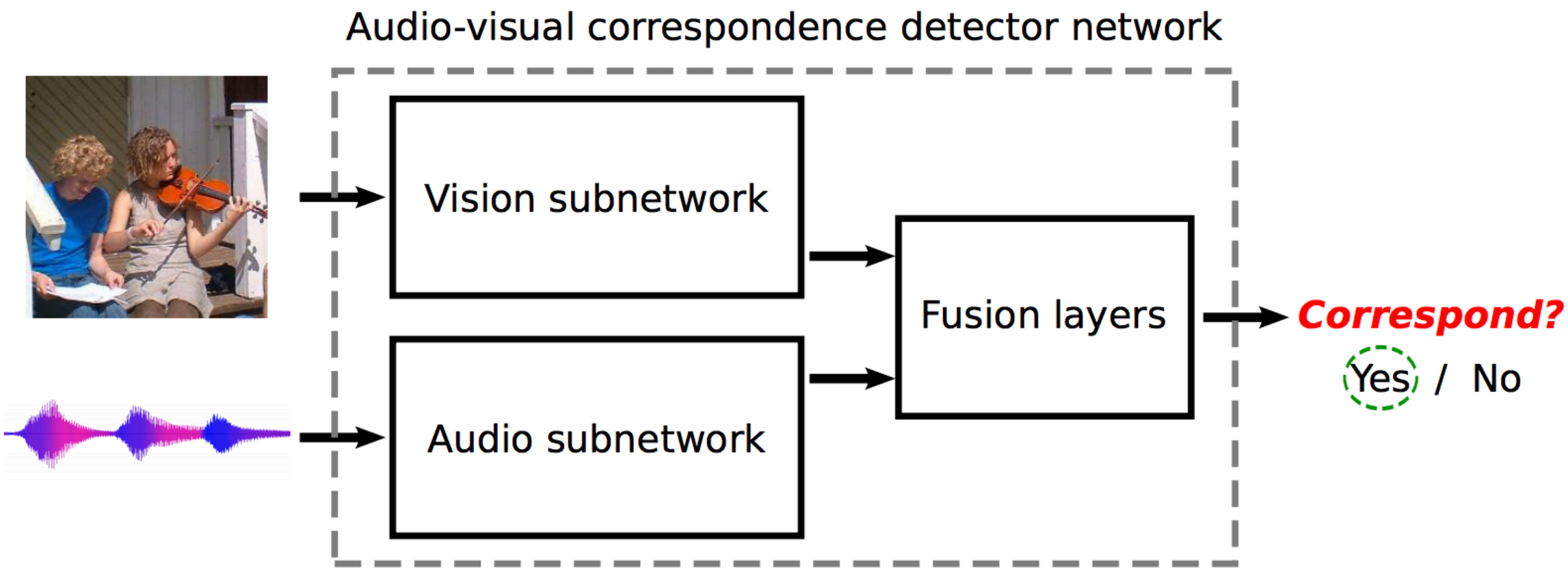
Idea #1: random pairs



,

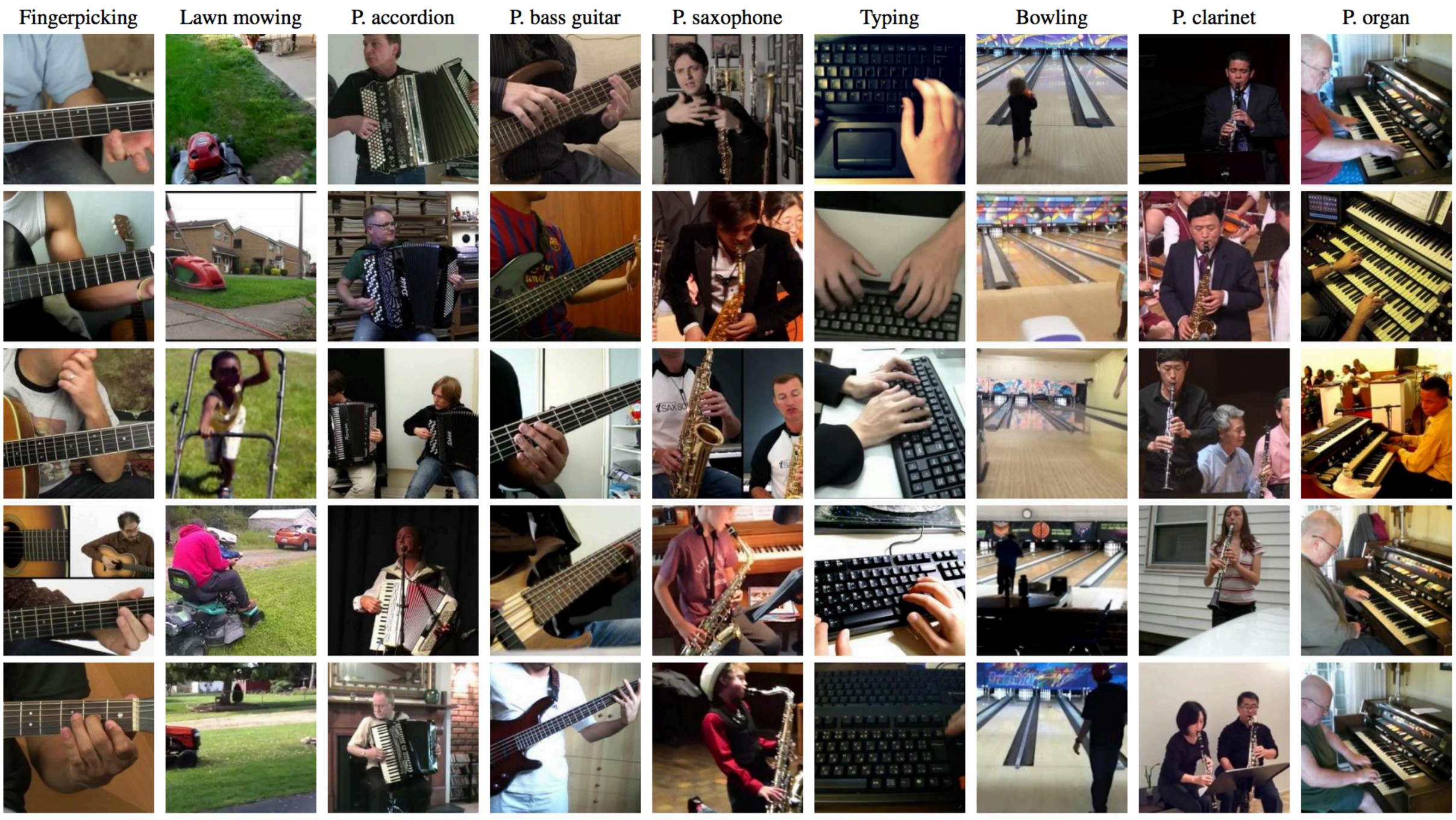
Arandjelovic, Zisserman. ICCV 2017

Slide credit: Andrew Owens



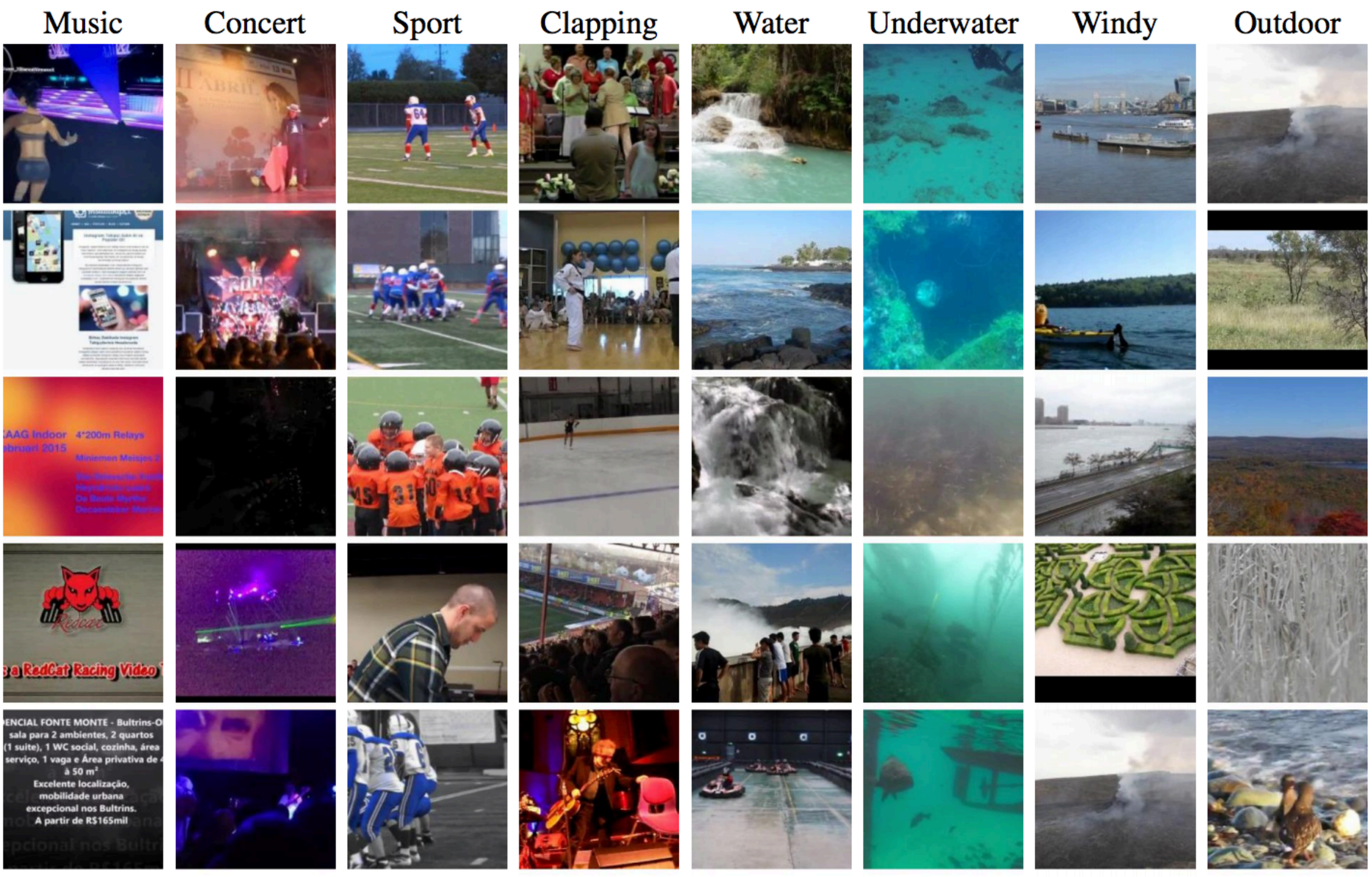
Arandjelovic, Zisserman. ICCV 2017

Vision hidden units



Arandjelovic, Zisserman. ICCV 2017

Sound hidden units



Arandjelovic, Zisserman. ICCV 2017

Sound Recognition

(a) ESC-50

Method	Accuracy
SVM-MFCC [26]	39.6%
Autoencoder [2]	39.9%
Random Forest [26]	44.3%
Piczak ConvNet [25]	64.5%
SoundNet [2]	74.2%
Ours random	62.5%
Ours	**79.3%**
Human perf. [26]	*81.3%*

(b) DCASE

Method	Accuracy
RG [27]	69%
LTT [19]	72%
RNH [28]	77%
Ensemble [32]	78%
SoundNet [2]	88%
Ours random	85%
Ours	**93%**

Arandjelovic, Zisserman. ICCV 2017

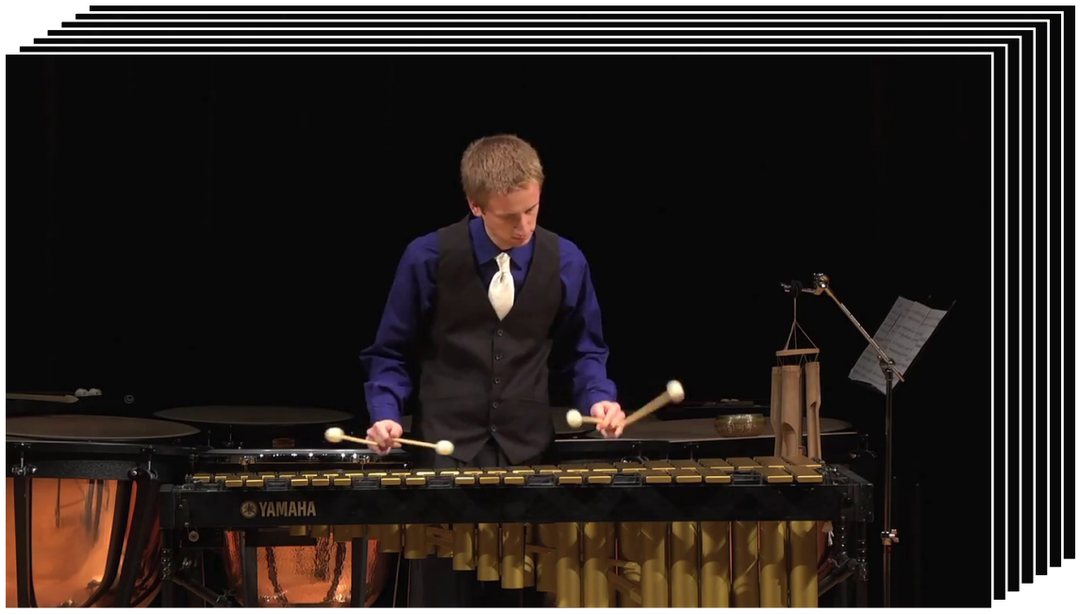
Visual Recognition

Method	Top 1 accuracy
Random	18.3%
Pathak *et al.* [24]	22.3%
Krähenbühl *et al.* [16]	24.5%
Donahue *et al.* [7]	31.0%
Doersch *et al.* [6]	31.7%
Zhang *et al.* [36] (init: [16])	32.6%
Noroozi and Favaro [21]	34.7%
Ours random	12.9%
Ours	32.3%

Linear classifier on top of features (ImageNet)

Arandjelovic, Zisserman. ICCV 2017

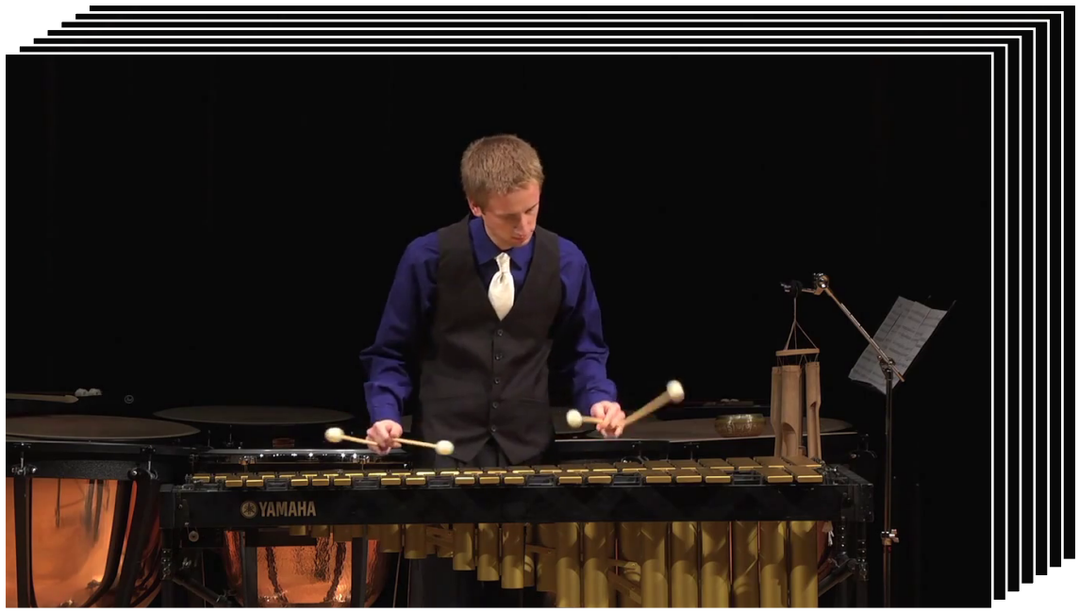
Idea #1: random pairs

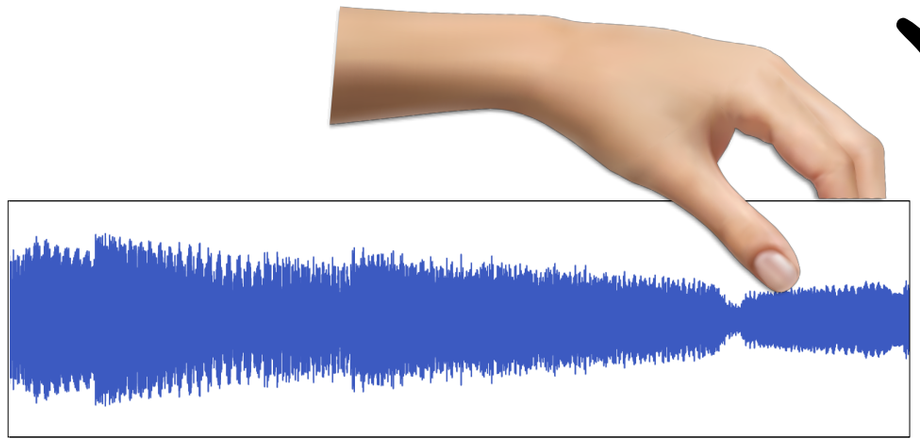


,

Slide credit: Andrew Owens

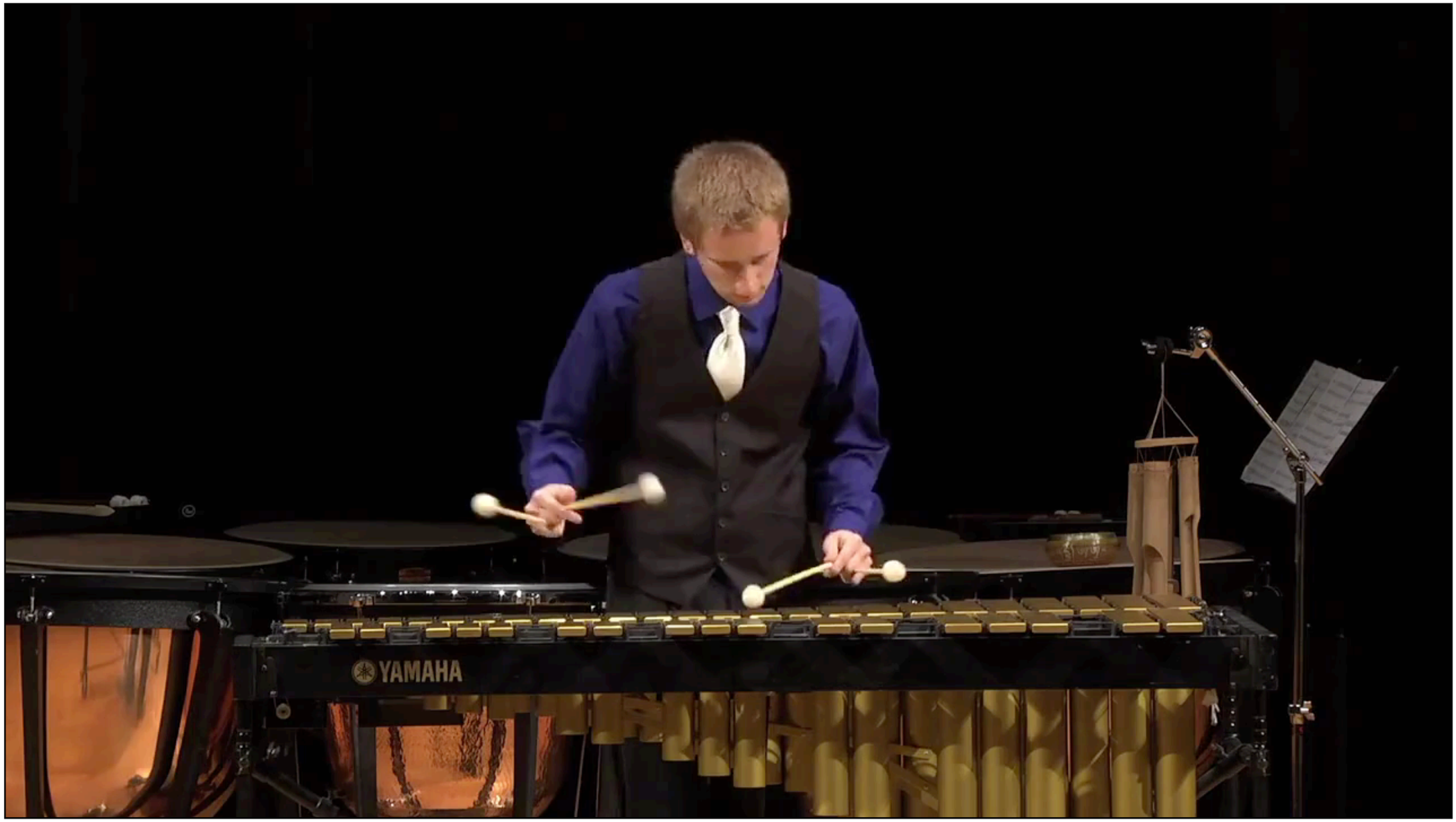
Idea #2: time-shifted pairs





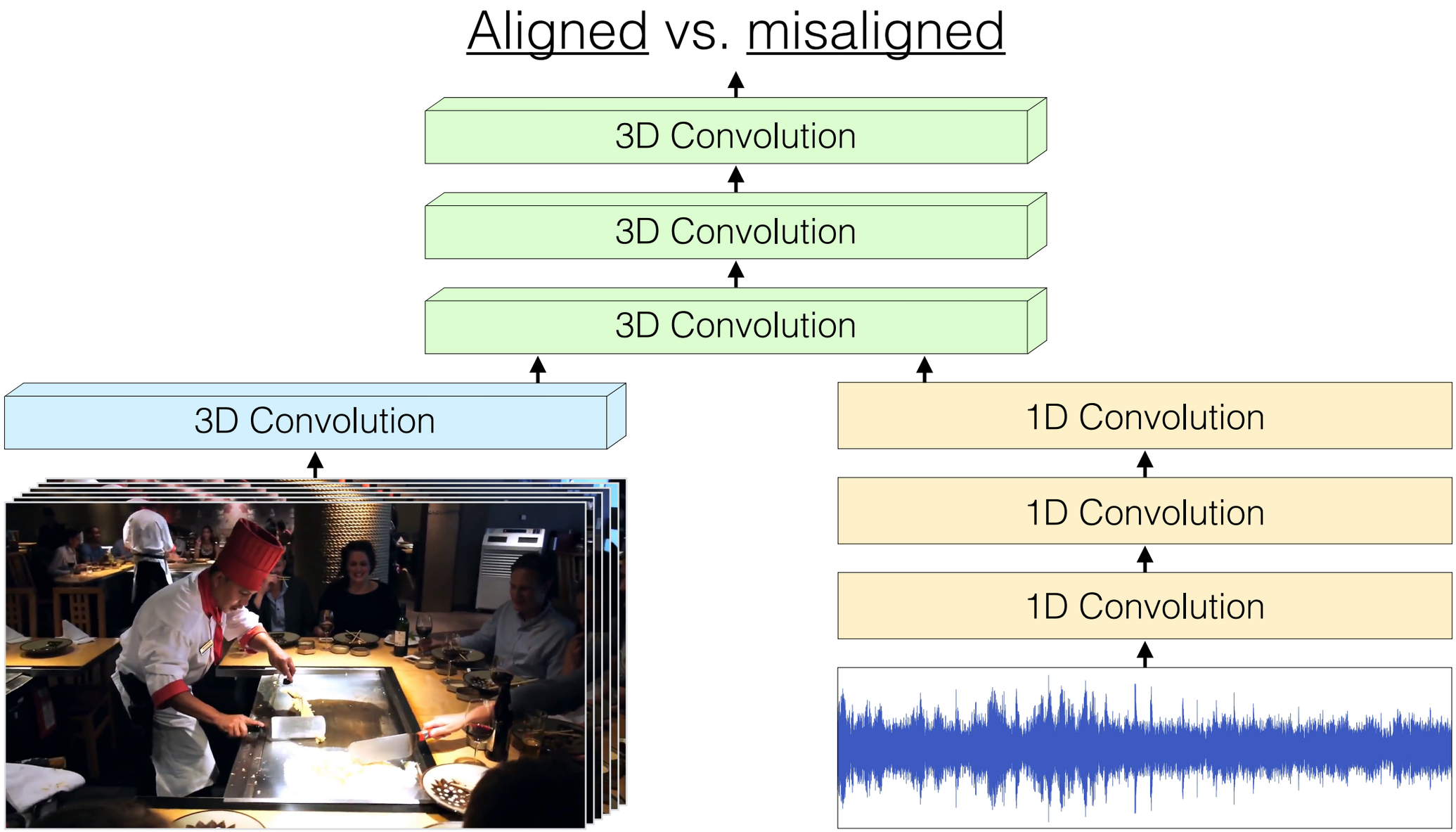
Slide credit: Andrew Owens

Idea #2: time-shifted pairs



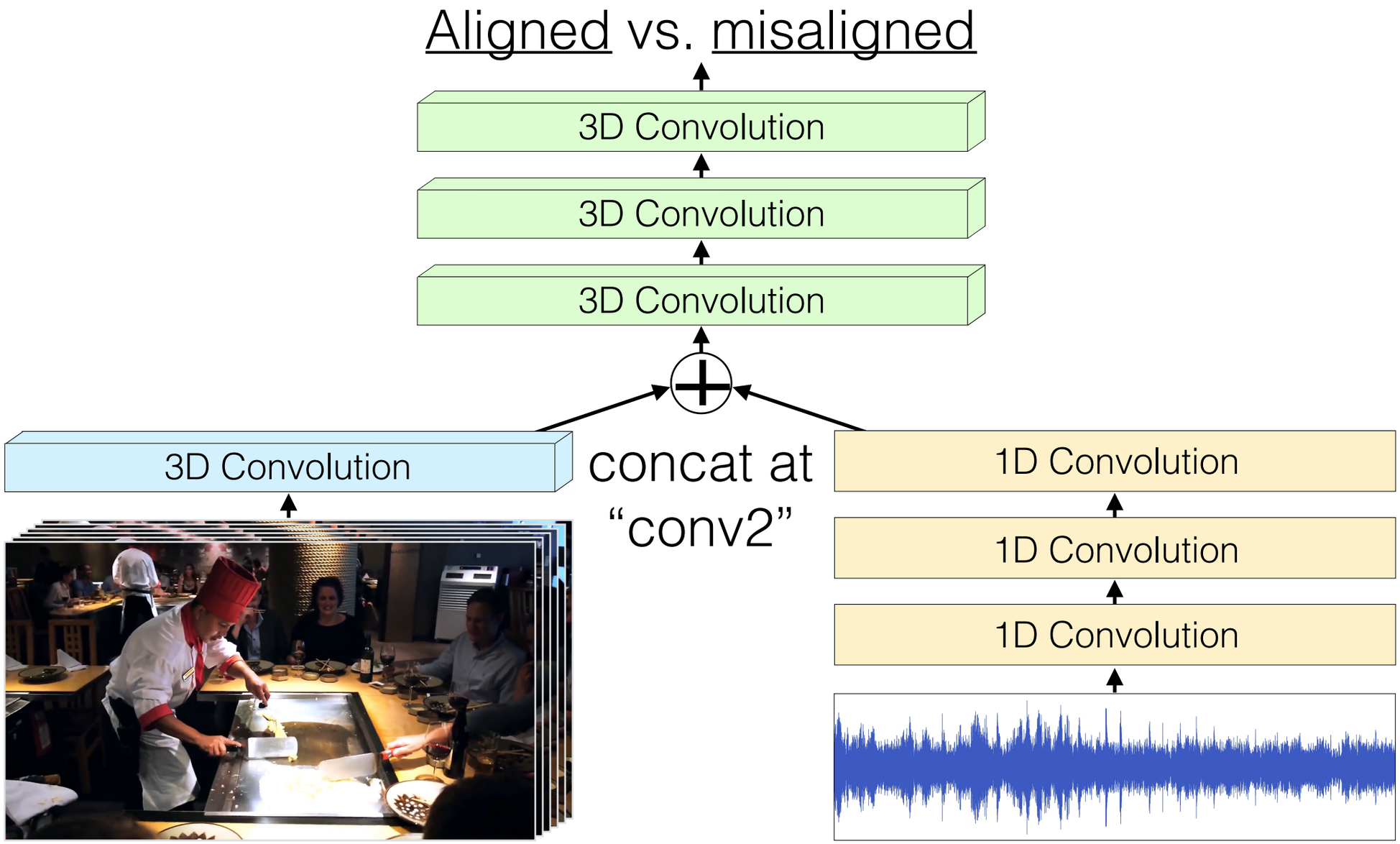
Slide credit: Andrew Owens

Fused audio-visual representation



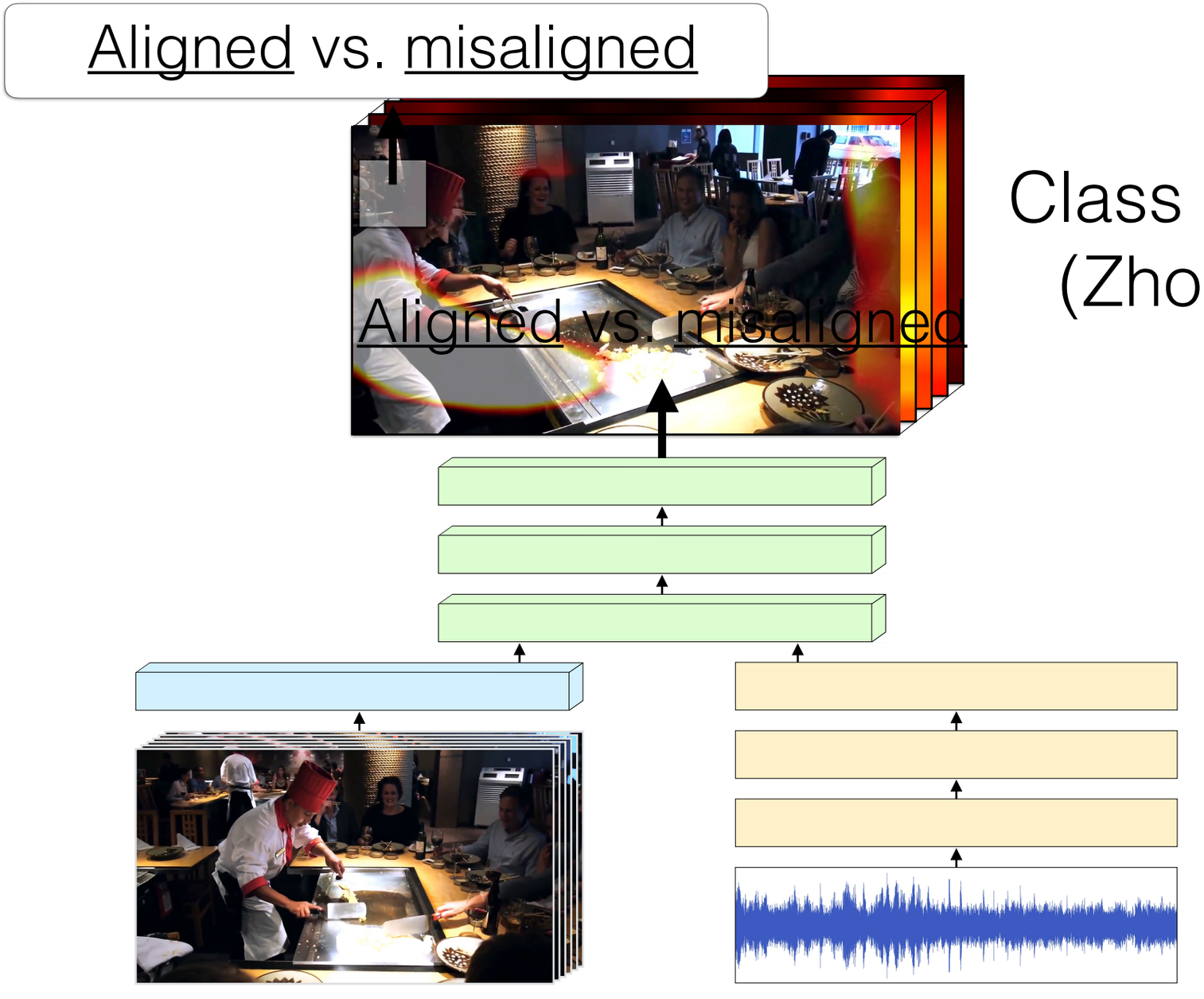
Slide credit: Andrew Owens

Fused audio-visual representation



Slide credit: Andrew Owens

What does the network learn?



Class activation map
(Zhou et al. 2016)

Slide credit: Andrew Owens

Top responses per category (speech examples omitted)

Dribbling basketball

Dribbling basketball

Dribbling basketball

Playing organ

Playing organ

Playing organ


Chopping wood

Chopping wood

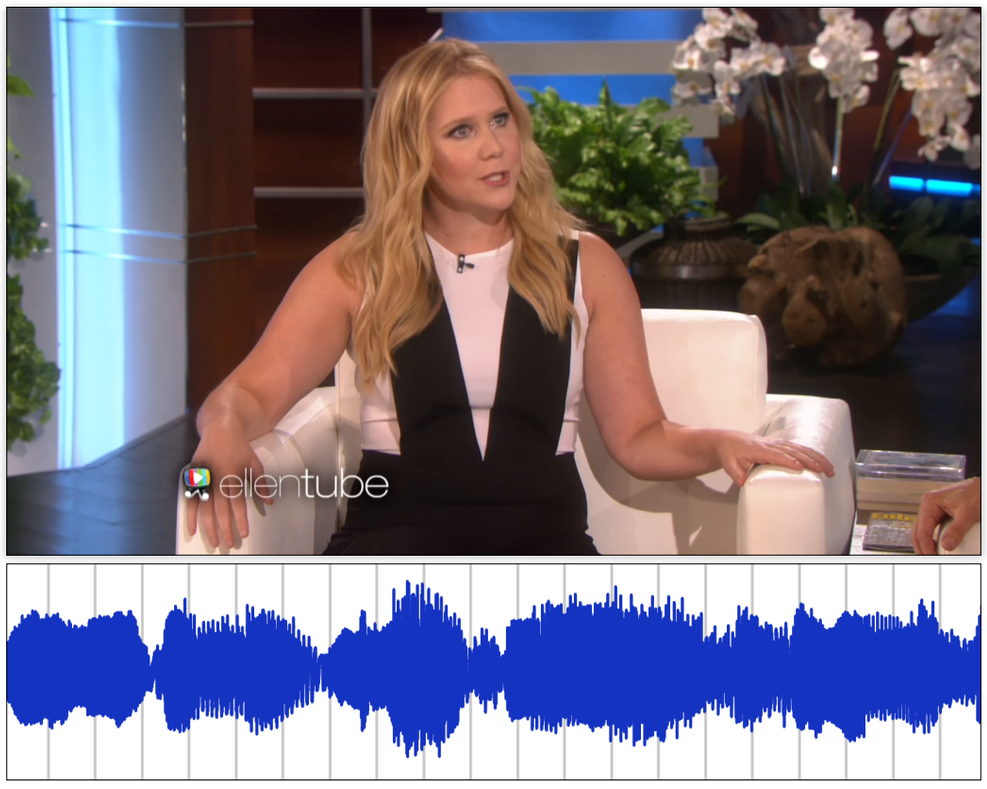
Chopping wood

Application: on/off-screen source separation

Task: separate on-screen sounds from background noise

Slide credit: Andrew Owens

Creating training data



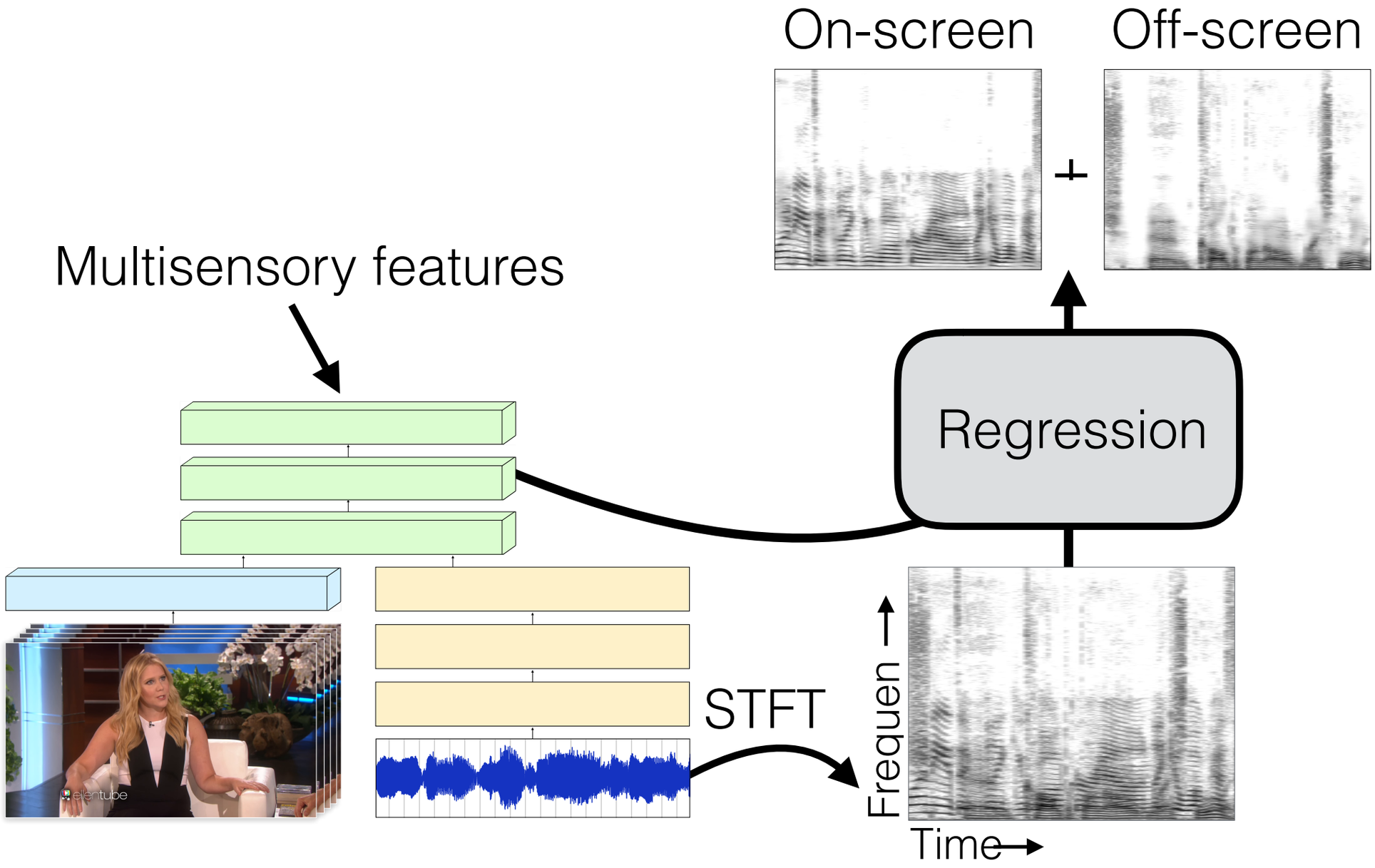
On-screen sound

Off-screen

Synthetic mixture

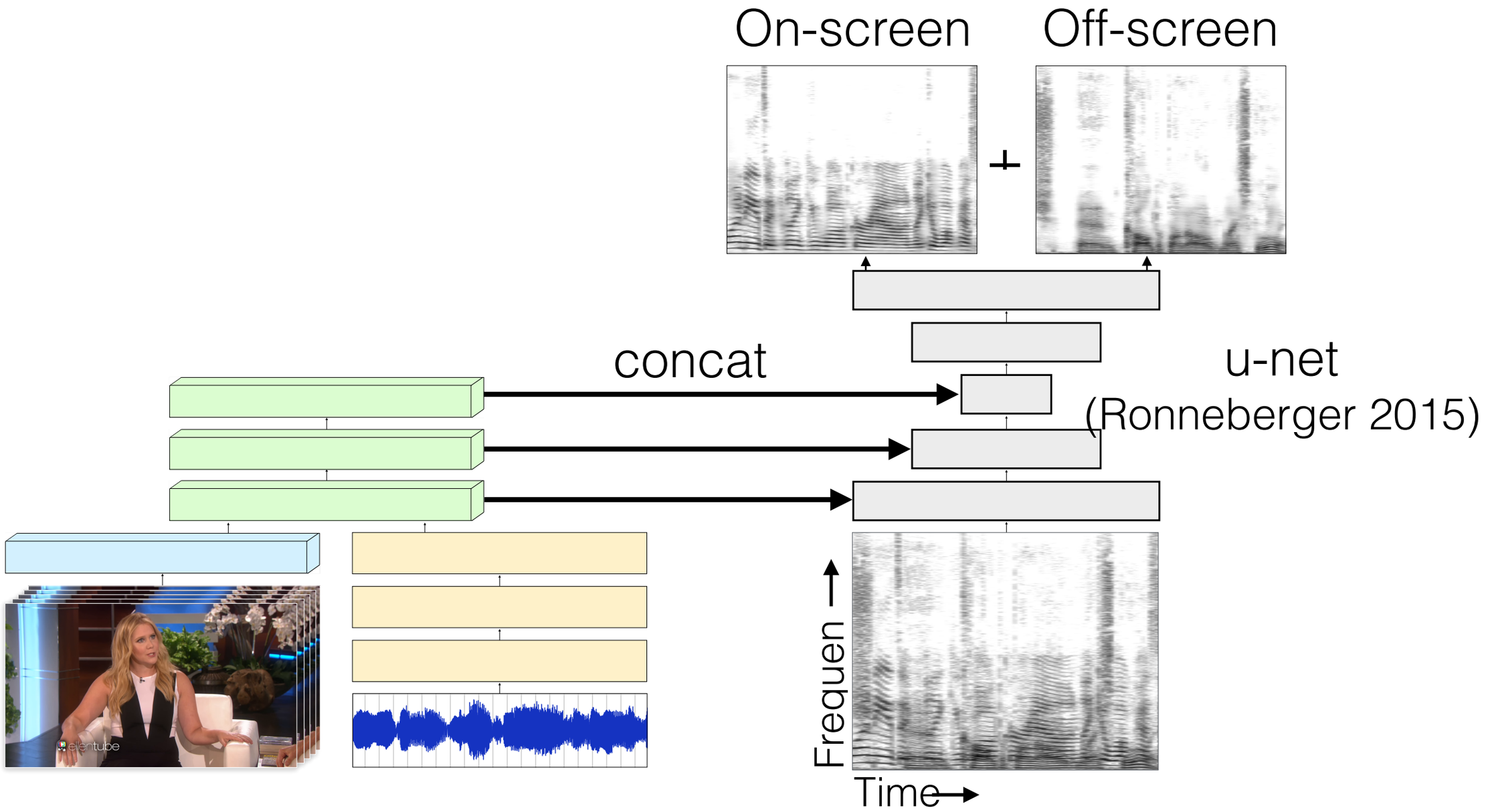
Slide credit: Andrew Owens

On/off-screen source separation



Slide credit: Andrew Owens

On/off-screen source separation

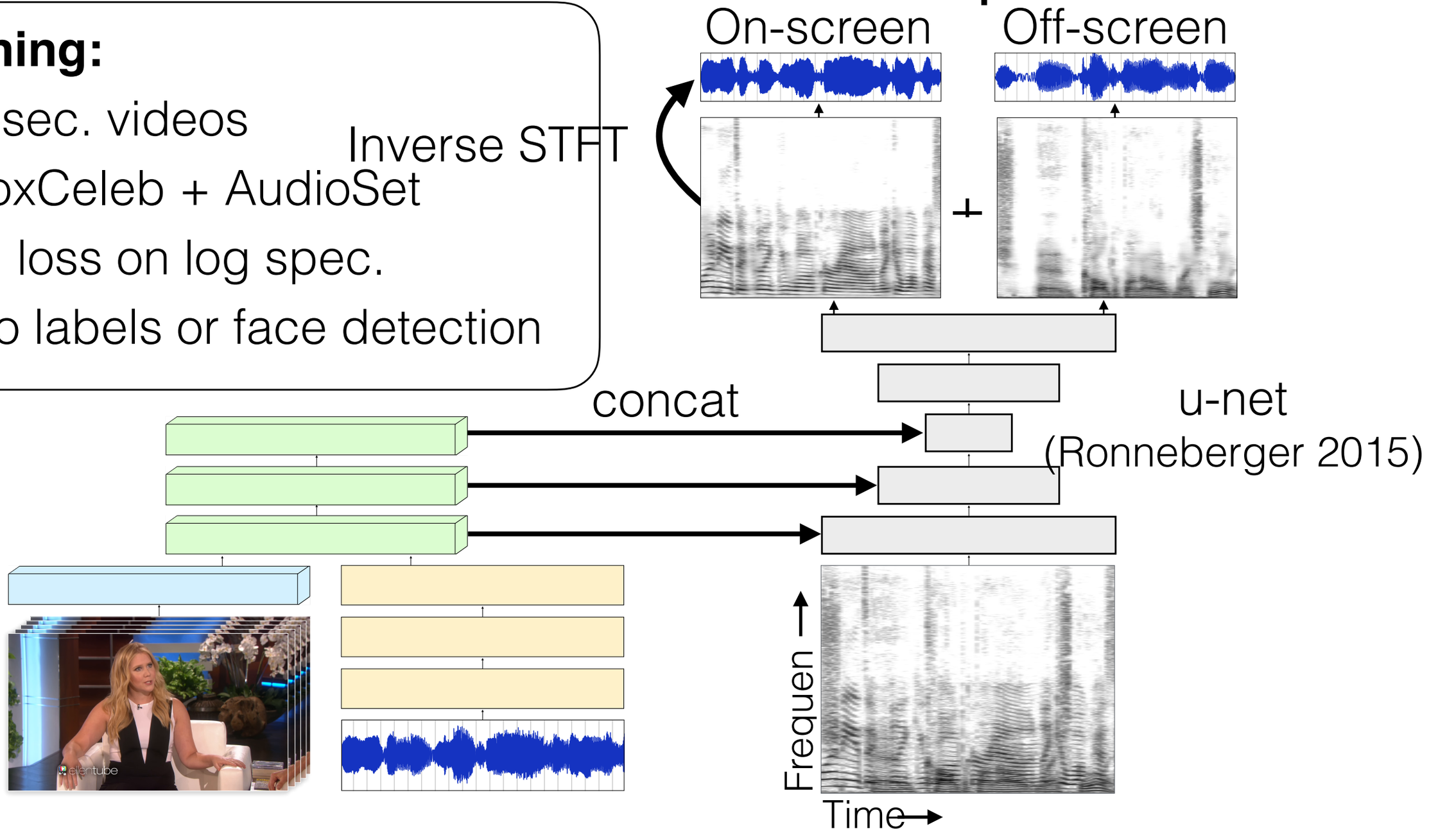


Slide credit: Andrew Owens

On/off-screen source separation

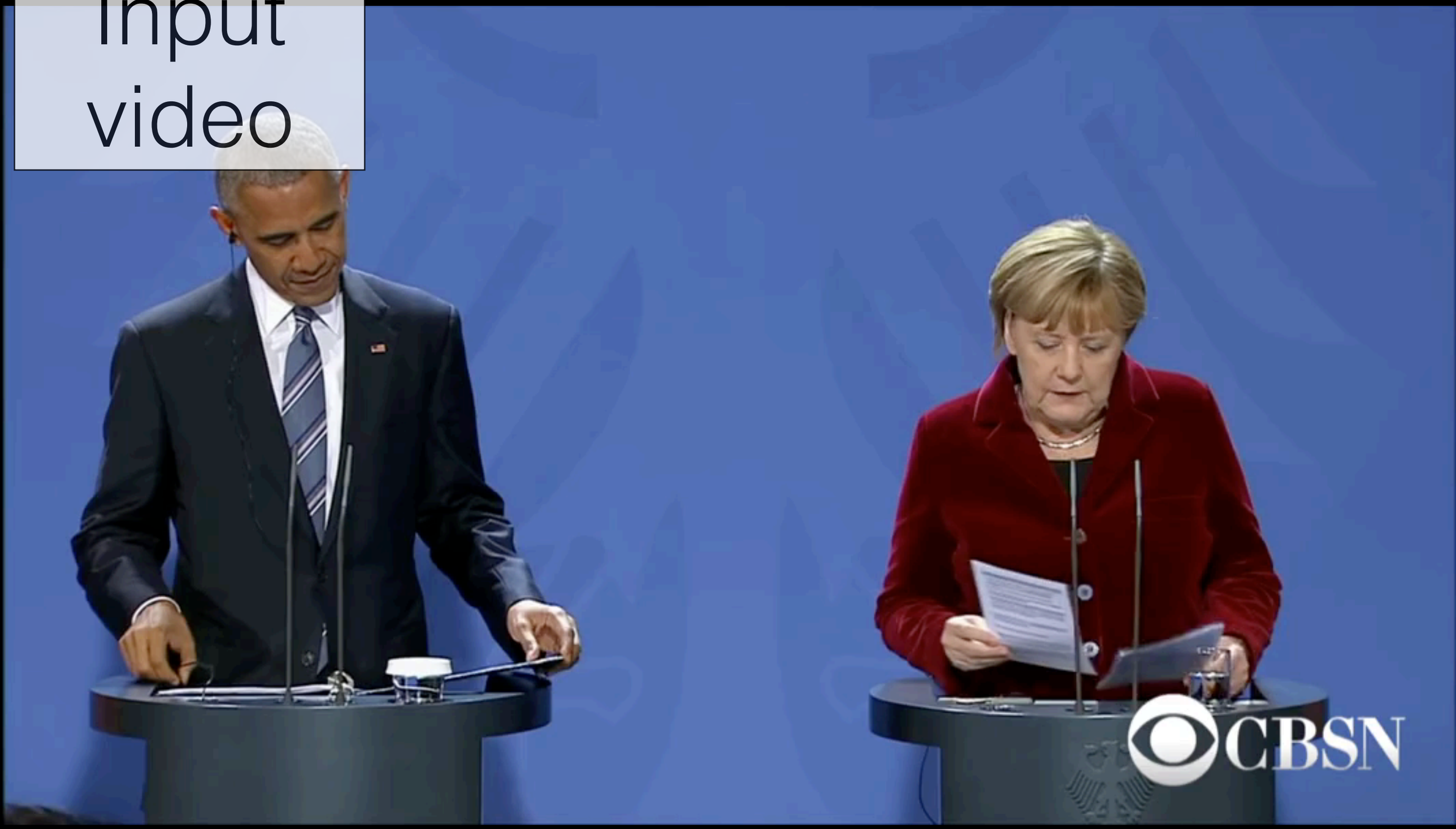
Training:

- 4-sec. videos
- VoxCeleb + AudioSet
- L_1 loss on log spec.
- No labels or face detection

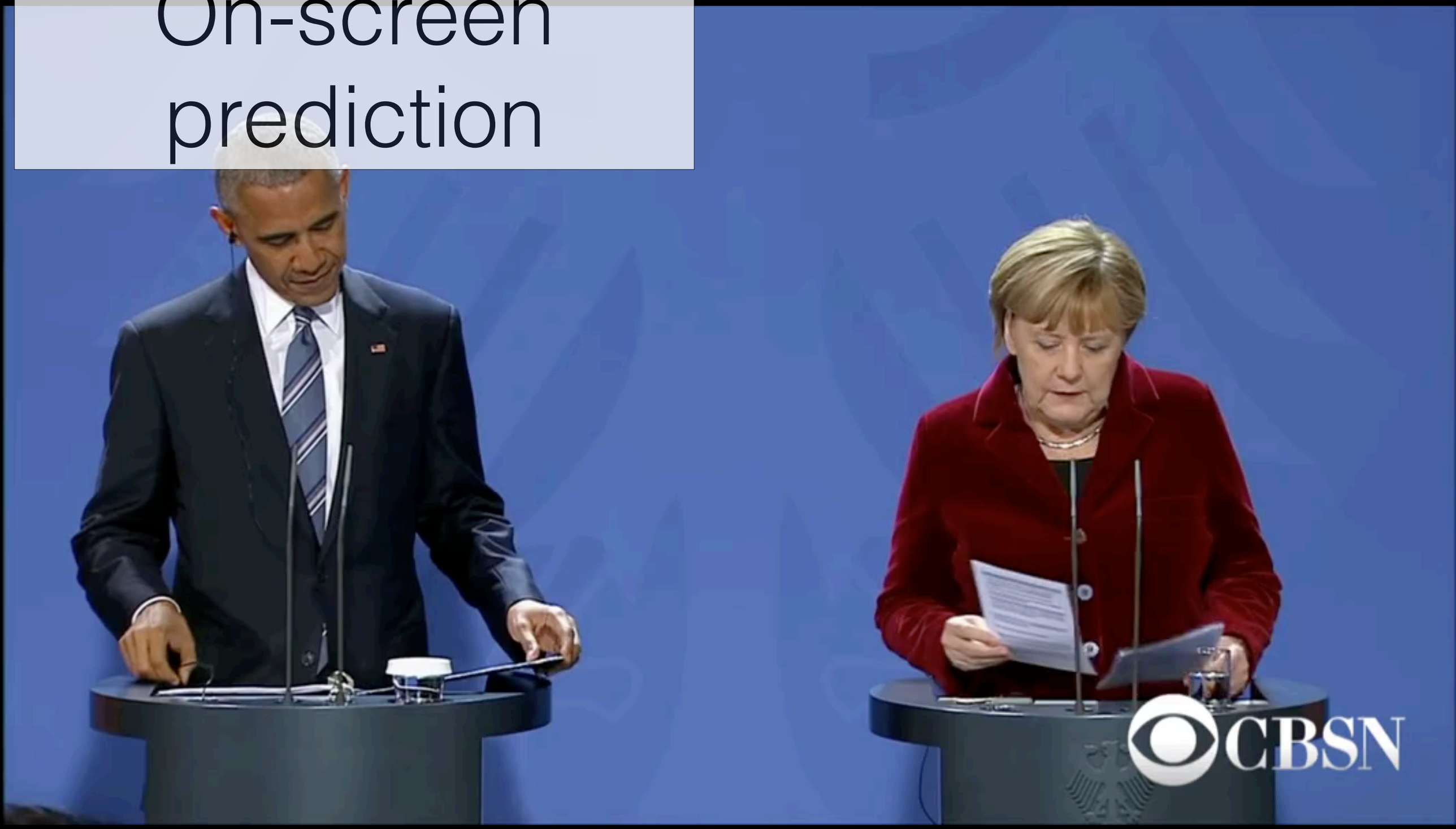


Slide credit: Andrew Owens

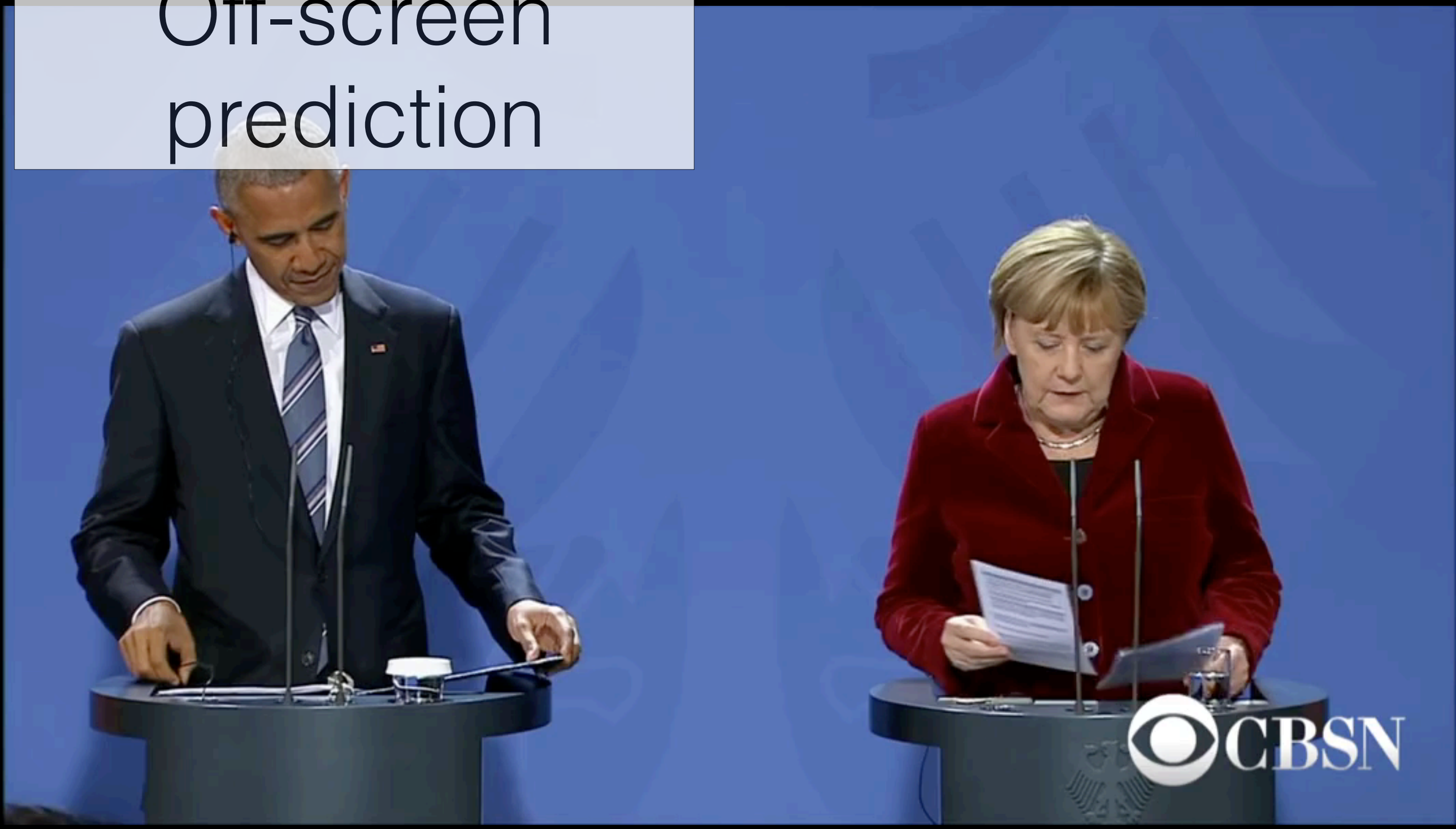
Input video



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