CIS Winter School, Dec. 2021

Separating Object Sounds from Videos and Robot Interactions

Ruohan Gao

Stanford University

Listening to learn about what we see

Listening to learn about what we see

Object identity

Material properties

Emotion

Dynamic sources

Ambient scene

Spatial cues

Sound of Objects

woof

meow

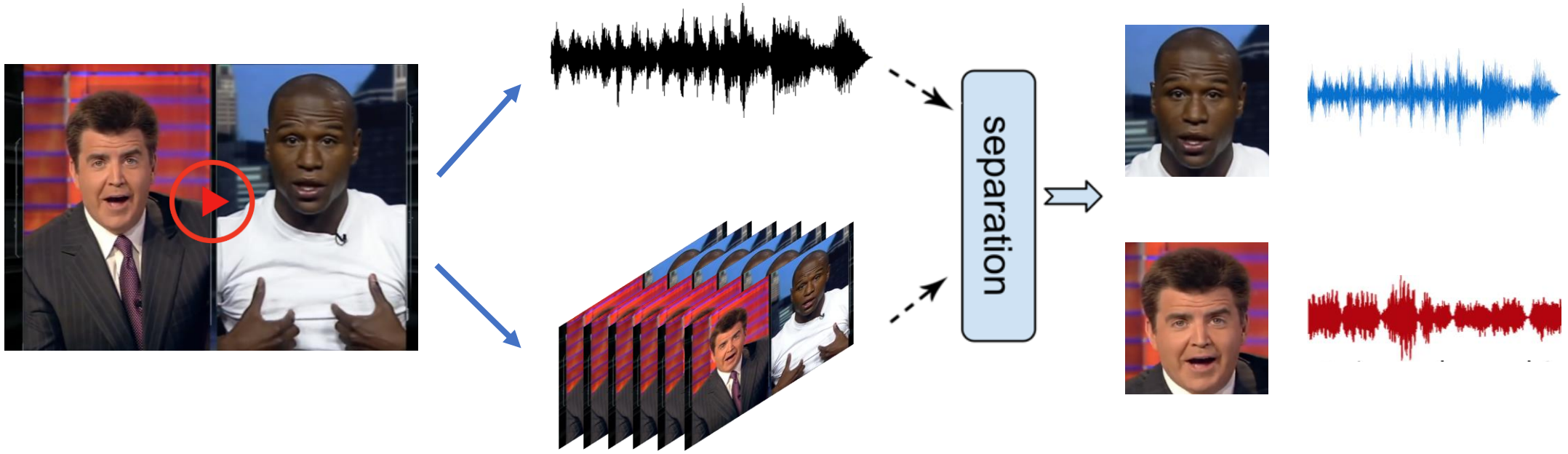
ring

clatter

Goal: a repertoire of objects and their sounds

Challenge: a single audio channel usually mixes sounds of multiple objects

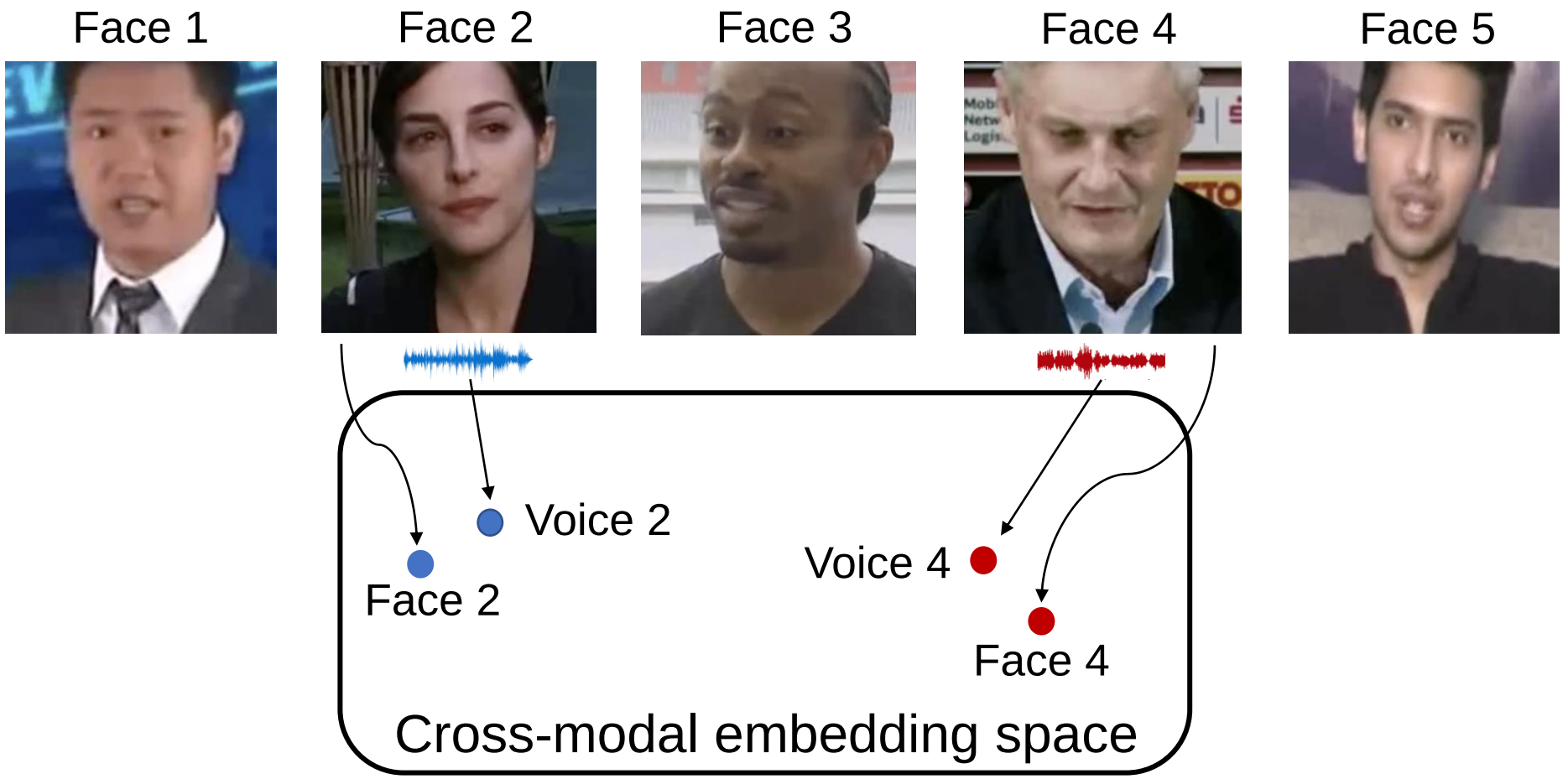
Visually-guided speech separation



Prior approaches analyze the facial motion (lip movements) in concert with the emitted speech.

[Afouras et al. Interspeech'18, Gabby et al. Interspeech'18, Owens & Efros ECCV'18, Ephrat et al. SIGGRAPH'18]

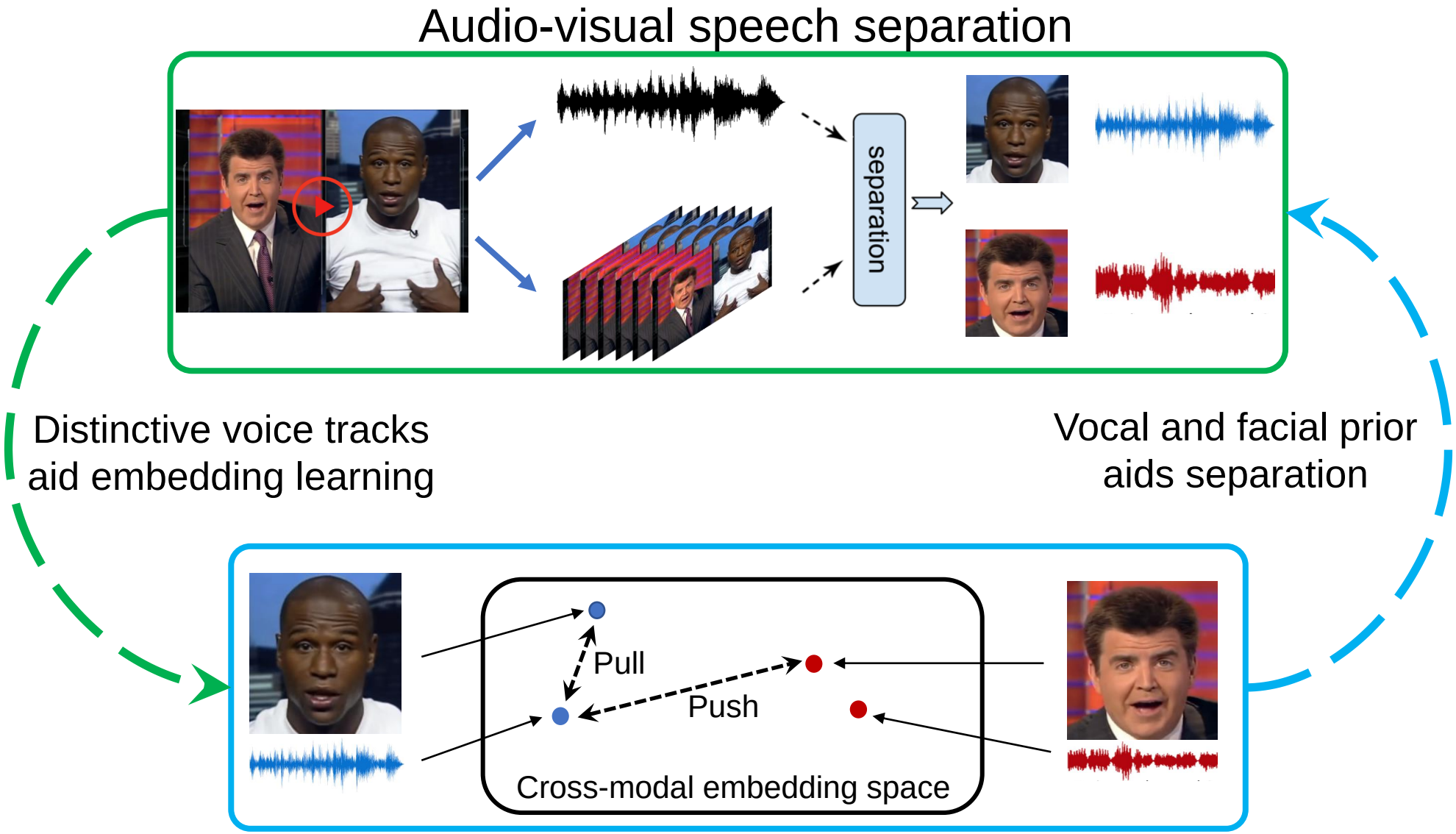
Facial appearance reveals voice qualities



Prior work on cross-modal matching learn cross-modal face-voice embeddings for the purpose of person identification.

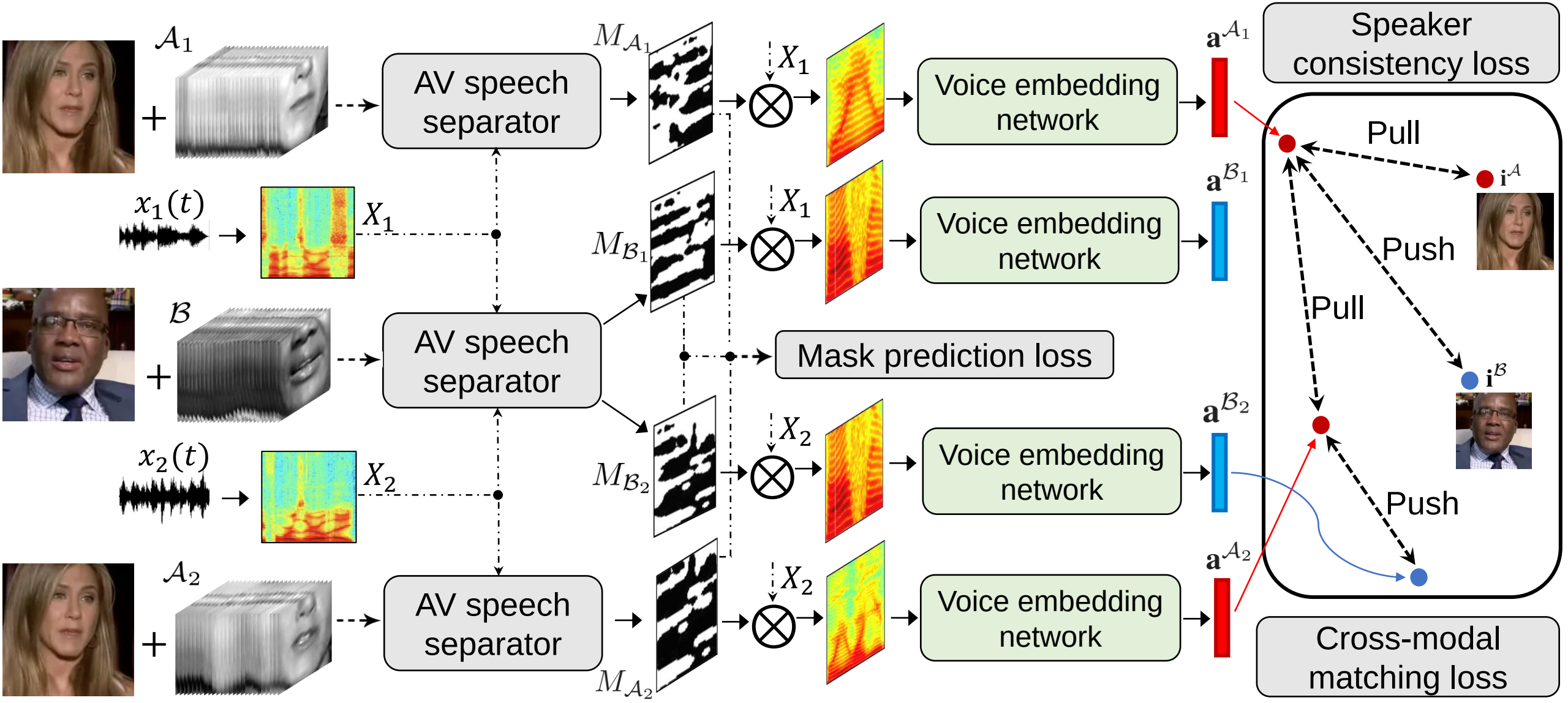
[Nagrani et al. ECCV'18, Nagrani et al. CVPR'18, Kim et al. ACCV'18, Chung et al. ICASSP'19, Wen et al. ICLR'19]

The two tasks are mutually beneficial



[Gao & Grauman, CVPR 2021]

Speech separation with cross-modal consistency



[Gao & Grauman, CVPR 2021]


Speech mixture


Separated voice for the left speaker


Separated voice for the right speaker


Speech with background noise


Enhanced speech

Results: Comparing to prior state-of-the-art methods

	Gabbay *et al.*	Hou *et al.*	Ephrat *et al.*	Ours
PESQ	2.25	2.42	2.50	**2.51**
STOI	–	0.66	0.71	**0.75**
SDR	–	2.80	6.10	**6.69**

(a) Results on Mandarin dataset.

	Gabbay *et al.*	Ephrat *et al.*	Ours
SDR	0.40	4.10	**10.9**
PESQ	2.03	2.42	**2.91**

(b) Results on TCD-TIMIT dataset.

	Casanovas *et al.*	Pu *et al.*	Ephrat *et al.*	Ours
SDR	7.0	6.2	12.6	**13.3**

(c) Results on CUAVE dataset.

	Afouras *et al.*	Afouras *et al.*	Ours
SDR	11.3	10.8	**11.8**
PESQ	3.0	3.0	**3.0**

(d) Results on LRS2 dataset.

	Chung *et al.*	Ours (static face)	Ours
SDR	2.53	7.21	**10.2**

(e) Results on VoxCeleb2 dataset.

Our method improves the state-of-the-art on all five datasets.

[Gao & Grauman, CVPR 2021]

Results

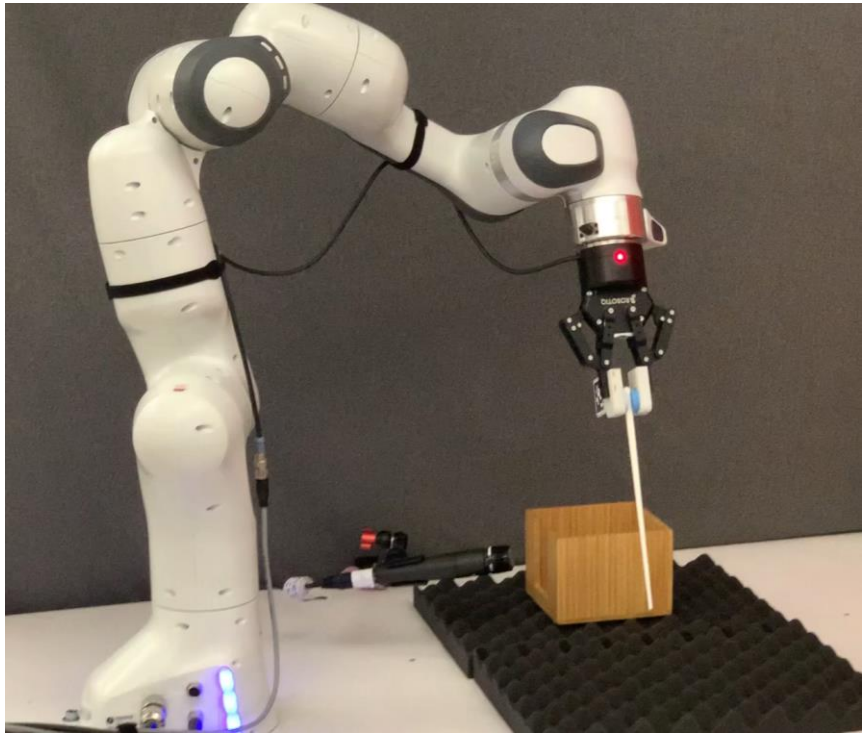
Train on 100,000 unlabeled multi-source video clips,
then separate audio for novel video.

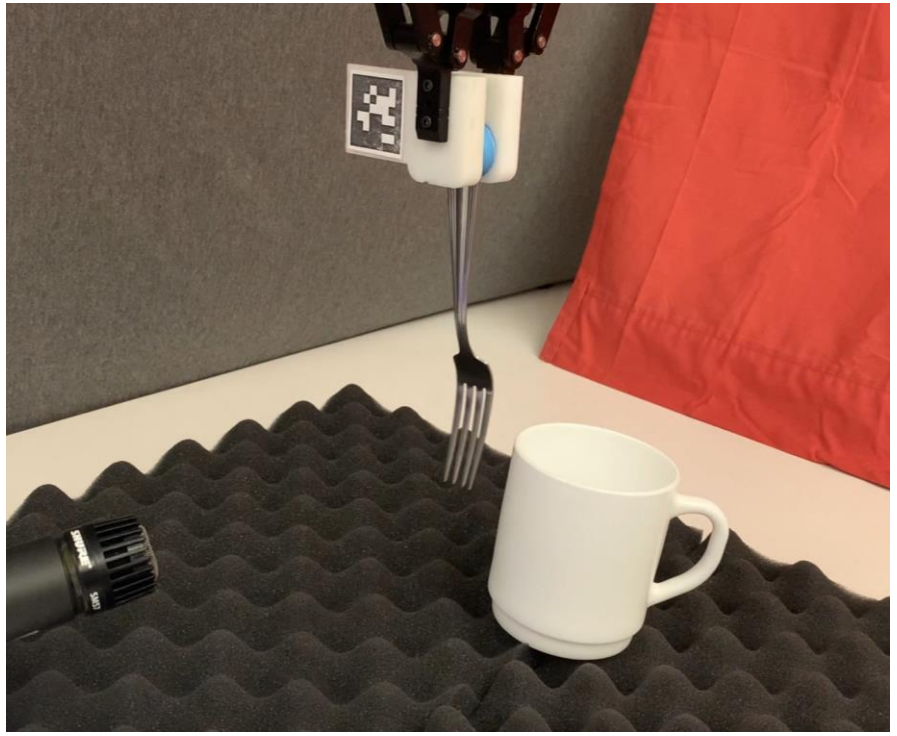
original video
(before separation)

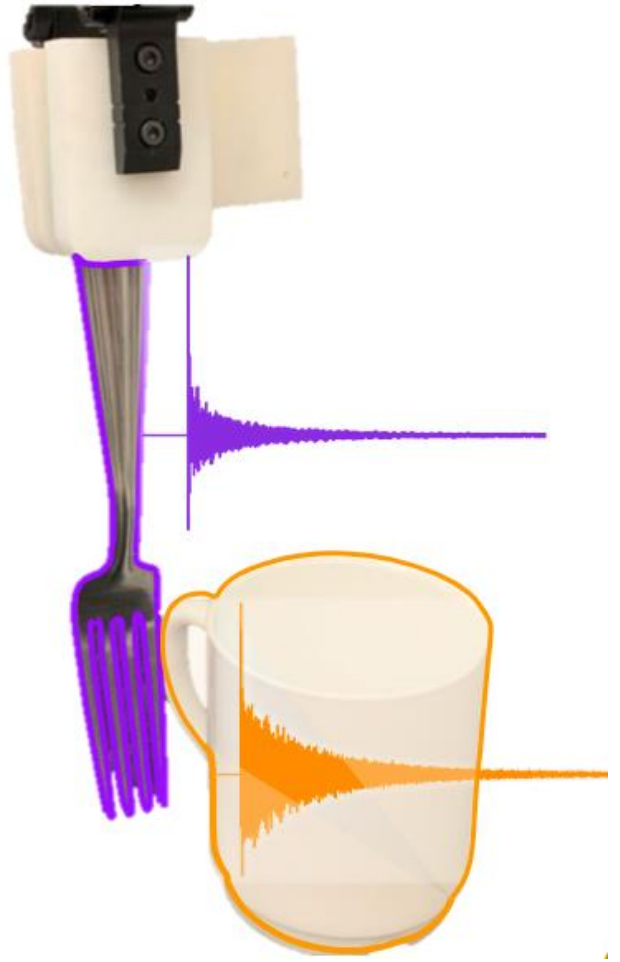
object detections:
violin & flute

[Gao et al. ECCV 2018, Gao & Grauman, ICCV 2019]

Separating Object Sounds for Robot Interactions

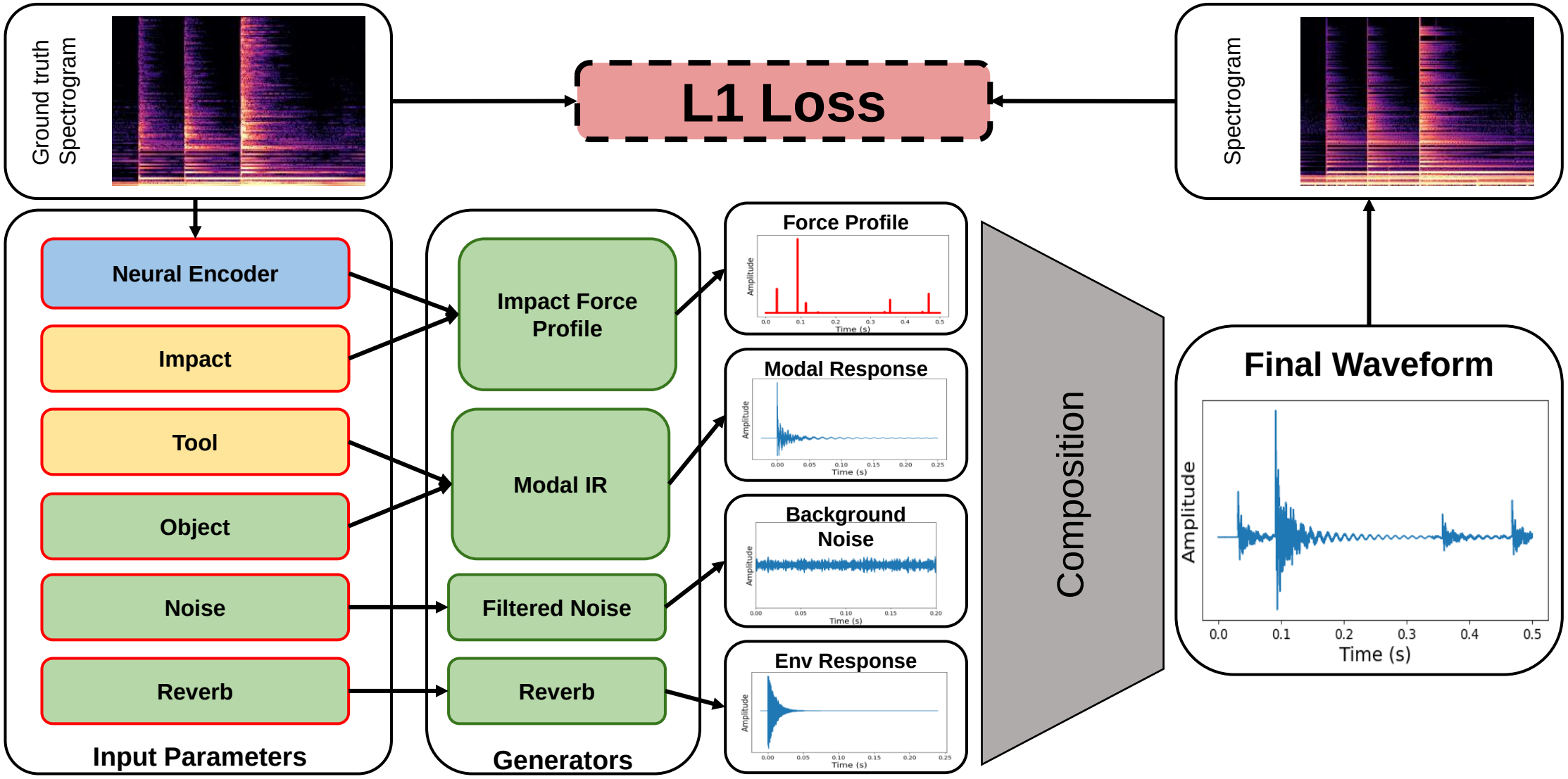




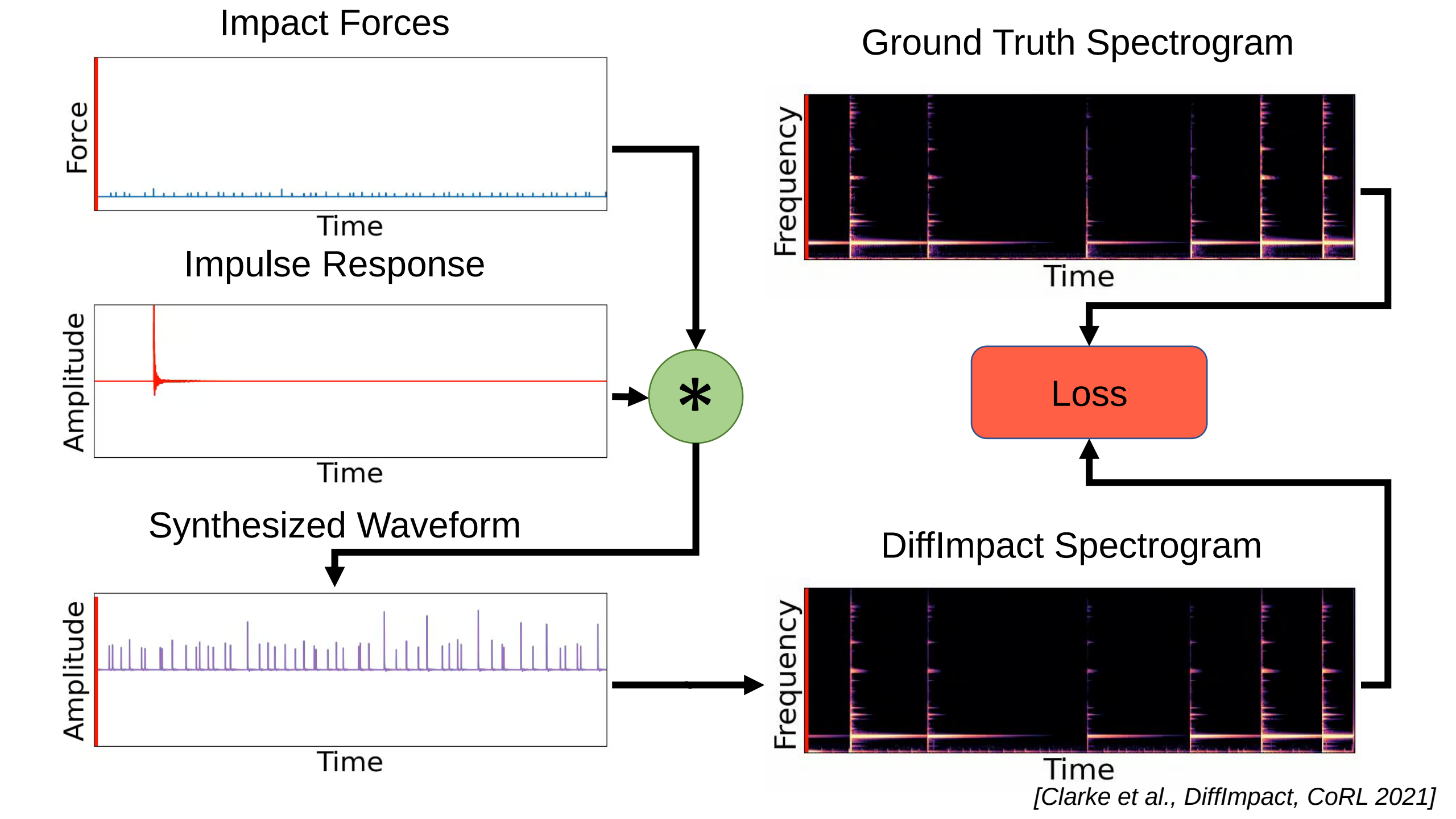


[Clarke et al., DiffImpact, CoRL 2021]

DiffImpact Model

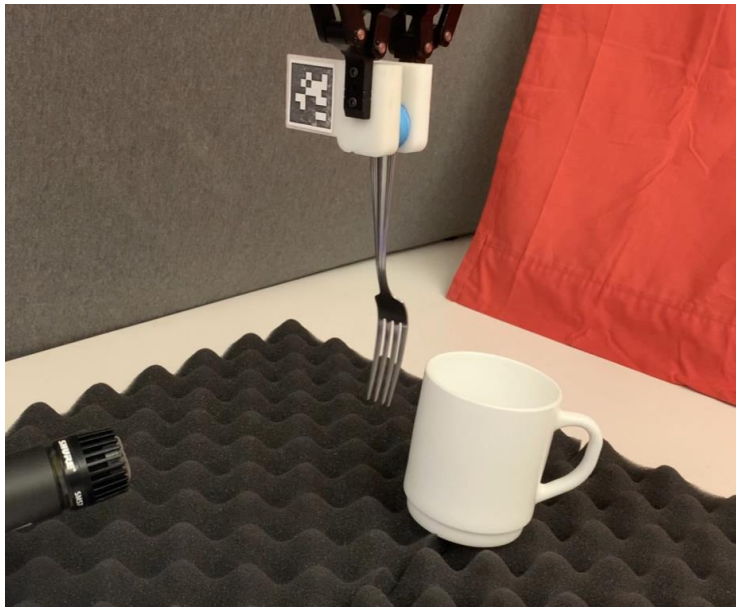


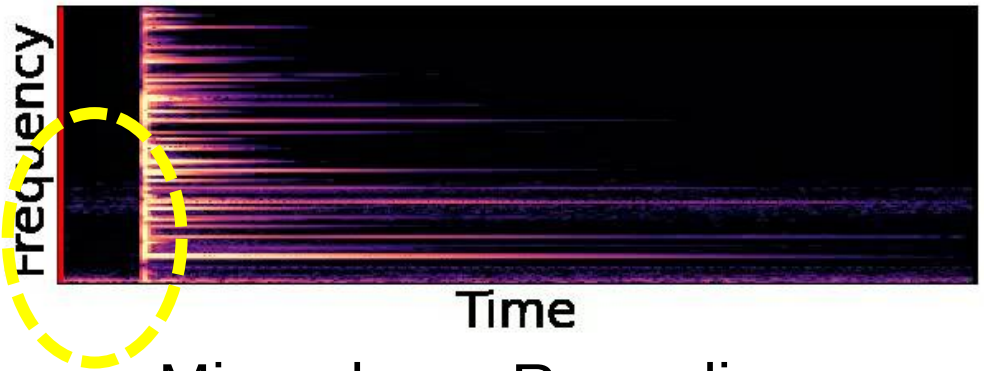
[Clarke et al., DiffImpact, CoRL 2021]



[Clarke et al., DiffImpact, CoRL 2021]

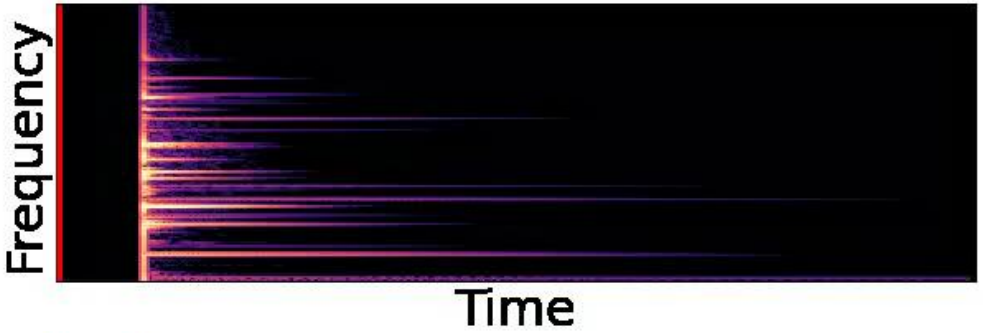
Separation of steel fork and and ceramic mug



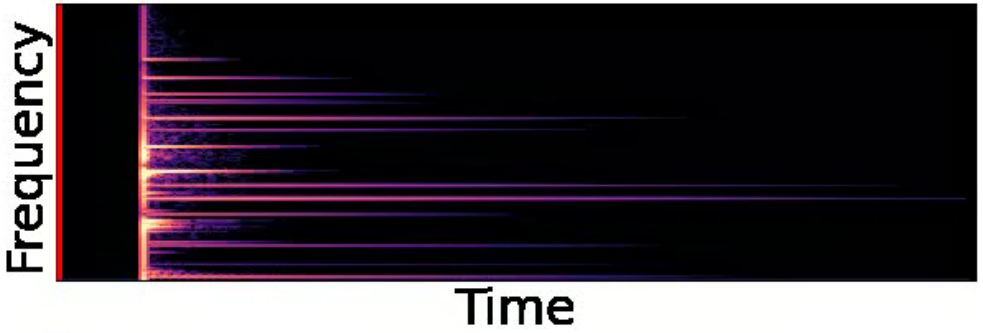


Microphone Recording

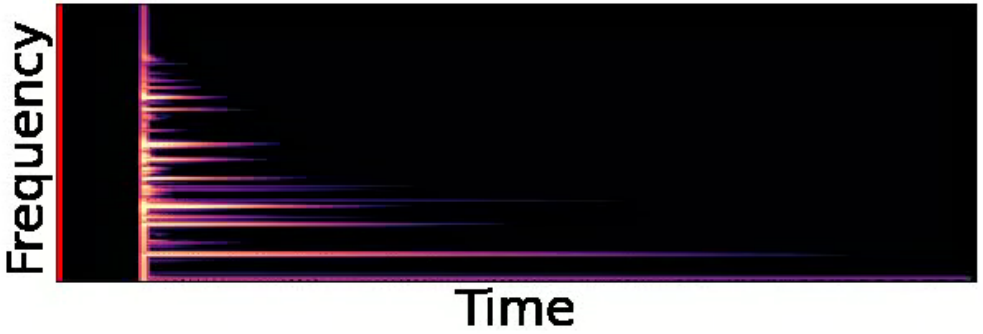
Synthesized Denoised



Separated Fork

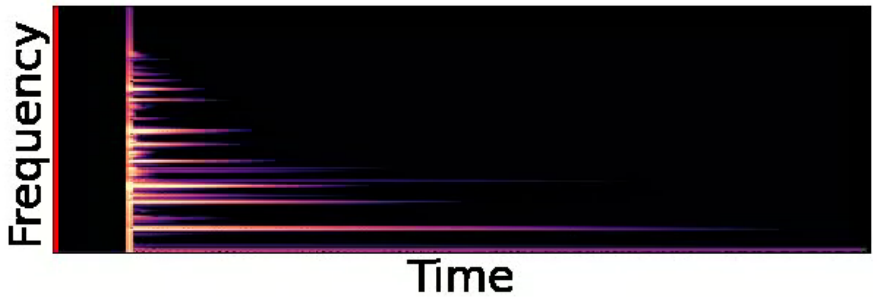


Separated Mug

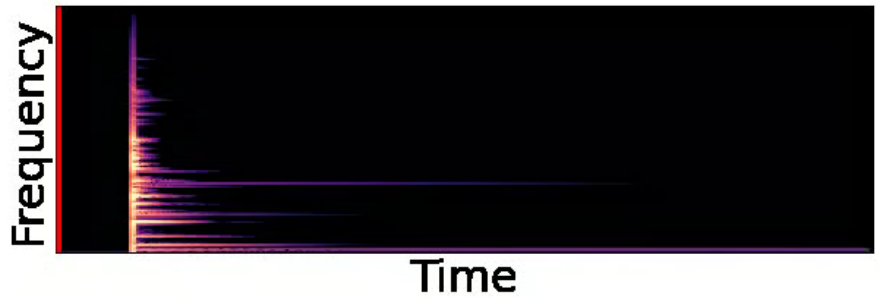


[Clarke et al., DiffImpact, CoRL 2021]

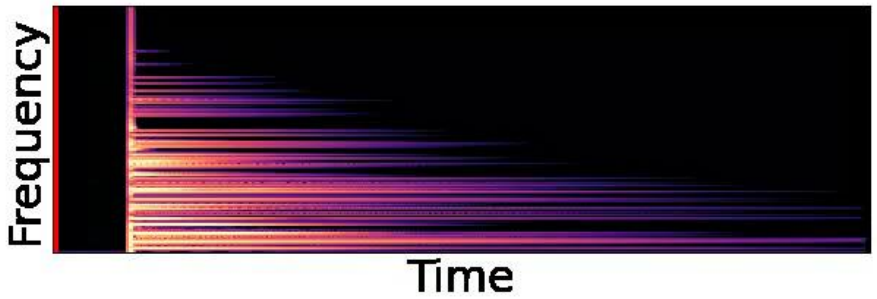
Separated Impacts from Each Object



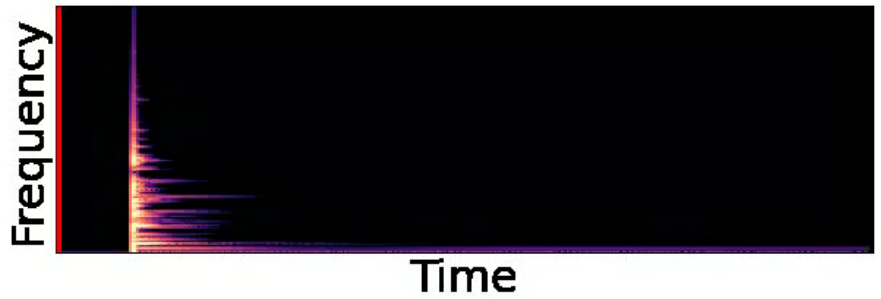
Ceramic Mug



Polycarbonate Cup



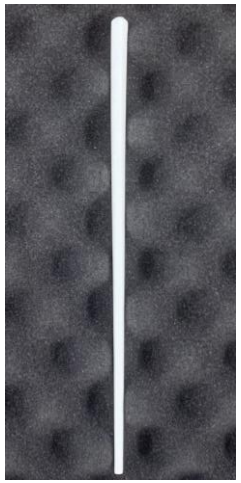
Steel Bowl

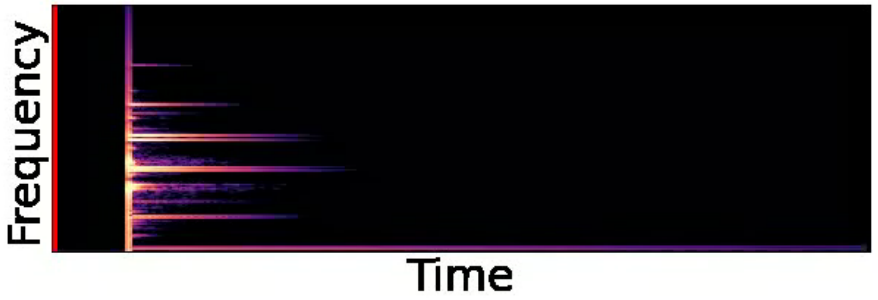


Wood Holder

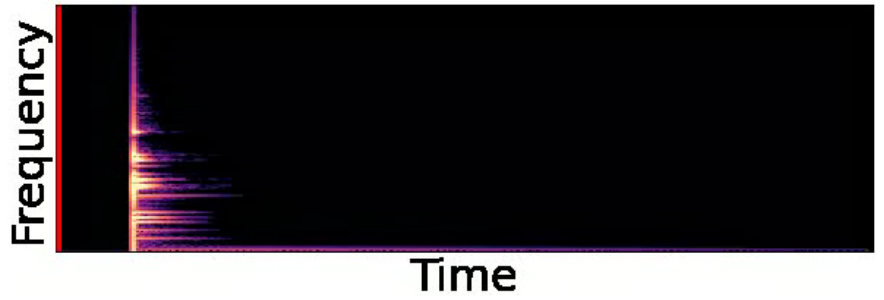
[Clarke et al., DiffImpact, CoRL 2021]

Separated Impacts from Each Tool

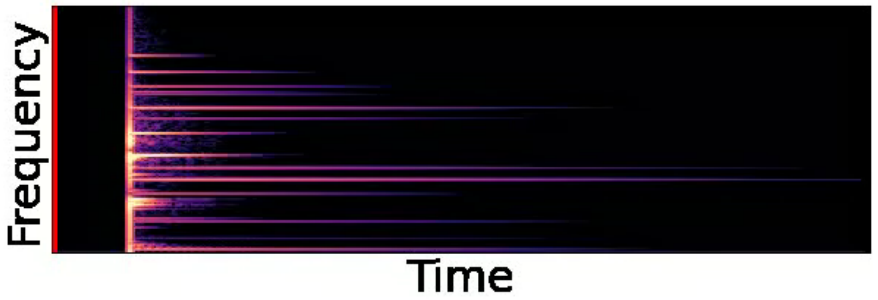




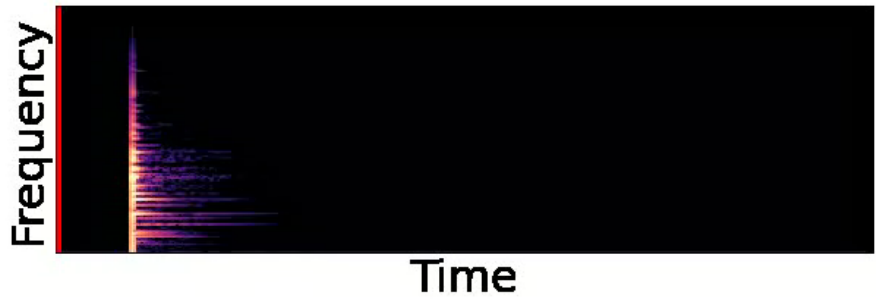
Ceramic Chopstick



Polycarbonate Spoon



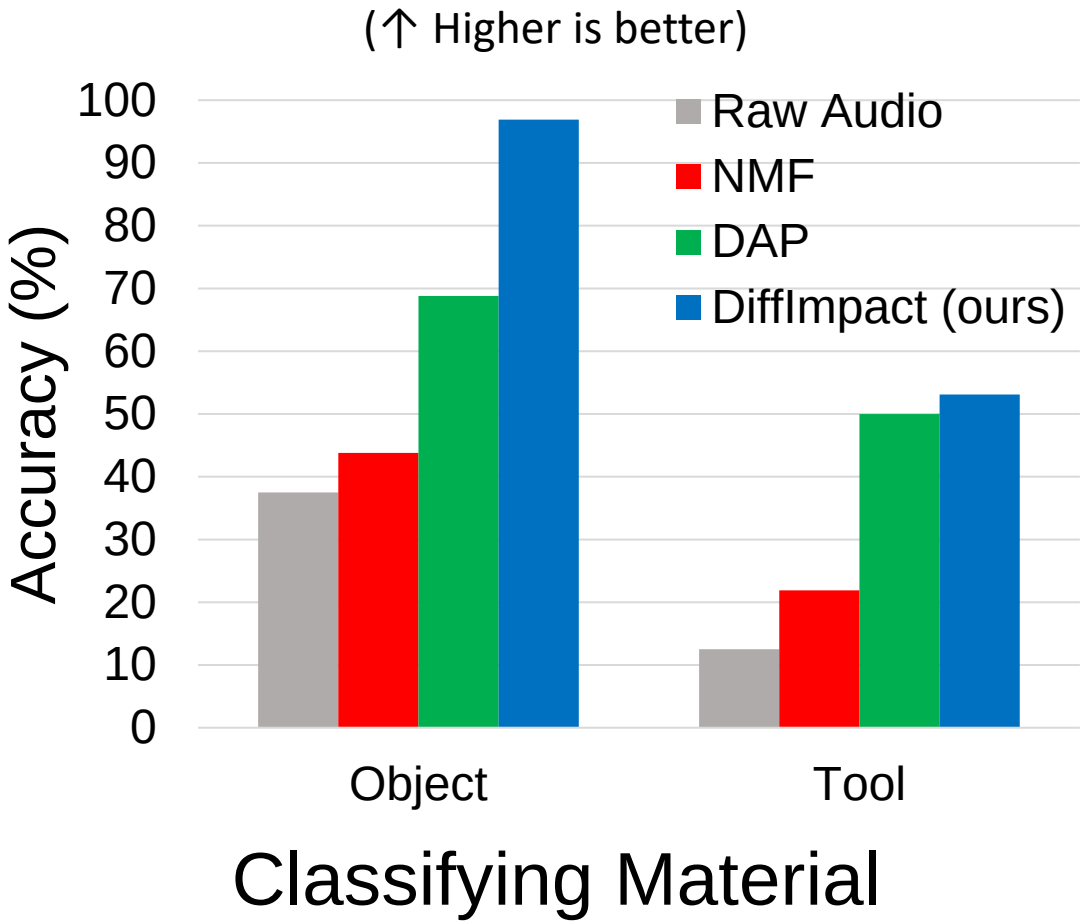
Steel Fork



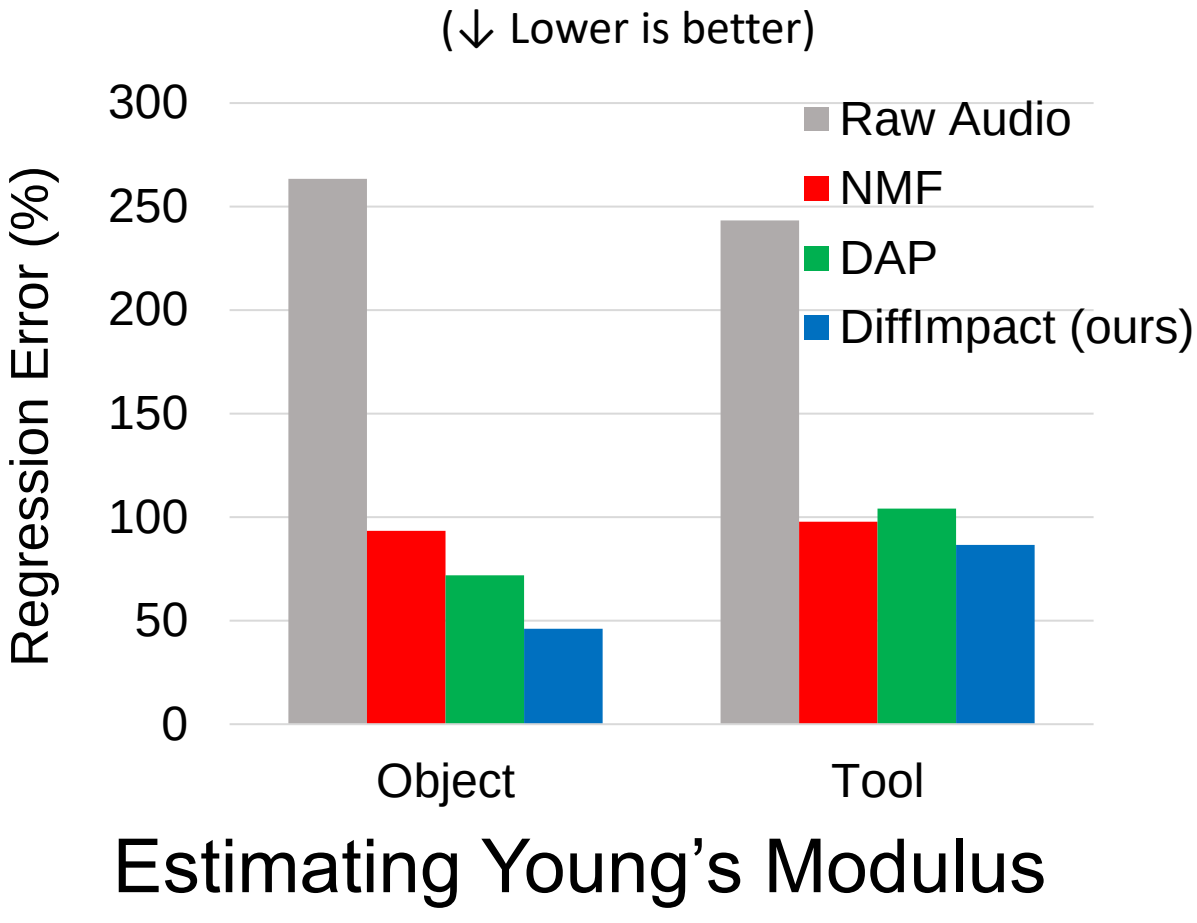
Wood Spoon

[Clarke et al., DiffImpact, CoRL 2021]

Inferring material properties of separated sounds



Classifying Material



Estimating Young's Modulus

NMF: Spiertz & Gnann, DAFX 2009
DAP: Tian et al. arXiv 2019

[Clarke et al., DiffImpact, CoRL 2021]

Summary

- Disentangling object sounds from videos
  - Visually-guided speech separation with cross-modal consistency (CVPR 2021)
- DiffImpact: A differentiable framework for rendering and identification of object-level impact sounds (CoRL 2021)

Kristen Grauman

Samuel Clarke

Jiajun Wu

Jeannette Bohg

Negin Heravi