SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, Dima Damen
Audio Signal – EPIC-KITCHENS

- Hand-object interactions

“crush bag”
Audio Signal – EPIC-KITCHENS

- Hand-object interactions
- Proximity of sensor to the ongoing action

“turn-on blender”
Audio Signal – EPIC-KITCHENS

- Hand-object interactions
- Proximity of sensor to the ongoing action
- Harmonic sounds

“rinse bell pepper”
Audio Signal – EPIC-KITCHENS

- Hand-object interactions
- Proximity of sensor to the ongoing action
- Harmonic sounds
- Percussive sounds

“chop garlic cloves”
Audio Signal – VGG-Sound

- Harmonic sounds

“thunder”
Audio Signal – VGG-Sound

• Harmonic sounds

“canary calling”
Audio Signal – VGG-Sound

- Harmonic sounds
- Percussive sounds

“typing on typewriter”
Audio Signal – VGG-Sound

- Harmonic sounds
- Percussive sounds

“playing tennis”
Two-stream motivation

• Strong evidence in neuroscience about ventral-dorsal streams in human auditory system
  - Some works suggest that ventral has high spectral resolution, while dorsal has high temporal resolution and operates at a higher sampling rate [1].

• Inspired by visual Slow-Fast net [2]

[1] Santoro et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLOS Computational Biology
[2] Feichtenhofer et al. (2019). SlowFast Networks for Video Recognition. International Conference on Computer Vision (ICCV)
Auditory Slow-Fast Network

Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

* 2D temporal convolution with kernel k x 1 and stride α
Auditory Slow-Fast Network
Auditory Slow-Fast Network

- Slow has low temporal precision and large amount of channels

Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Auditory Slow-Fast Network

- Slow has low temporal precision and large amount of channels
- Fast has fewer channels but high temporal resolution

Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Auditory Slow-Fast Network

- Slow has low temporal precision and large amount of channels
- Fast has fewer channels but high temporal resolution
- Multi-level lateral connections

Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Auditory Slow-Fast Network

- Slow has low temporal precision and large amount of channels
- Fast has fewer channels but high temporal resolution
- Multi-level lateral connections
- Separable convolutions

Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Results: VGG-Sound

| Model                        | Top-1  | Top-5  |
|------------------------------|--------|--------|
| Chen et al. [2]              | 51.00  | 76.40  |
| McDonnell & Gao [3]          | 39.74  | 71.65  |
| Slow                         | 45.20  | 72.53  |
| Fast                         | 41.44  | 70.68  |
| Slow-Fast (Proposed)         | 52.46  | 78.12  |
## Results: EPIC-KITCHENS

| Split | Model                  | Verb | Noun | Action | # Param. |
|-------|------------------------|------|------|--------|----------|
| Test  | Damen et al. [1]       | 42.12| 21.51| 14.76  | 10.67M   |
|       | Slow-Fast (Proposed)   | 46.47| 22.77| 15.44  | 26.88M   |

[1] Damen et al. (2020). Rescaling Egocentric Vision, arXiv
### Class-wise performance on VGG-Sound

| Slow stream | Fast stream |
|-------------|-------------|
| Animals     | Percussive sounds |
| baltimore oriole calling | footsteps on snow |
| cheetah chirrup           | snake rattle   |
| zebra braying             | tap dancing    |
| dinosaurs bellowing       | car engine knocking |
| horse neighing            | woodpecker pecking tree |
| black capped chickadee calling | chopping wood |
| cat hissing              | people clapping |
| cuckoo bird calling       | lawn mowing    |
| mosquito buzzing          | typing on typewriter |
| bull bellowing            | opening or closing car doors |
| whale calling             | playing tennis |
| Scenes                   | Voices        |
| volcano explosion         | singing choir |
| playing lacrosse          | people cheering |
| hair dryer drying         | people crowd  |
| sea waves                 | child speech  |
| playing tympani           | cat purring   |
| blowtorch igniting        | dog barking   |
| opening/closing electric car | race car  |
| windows                  | singing bowl  |
| thunder                  | vacuum cleaner cleaning floors |
| electric blender running  | toilet flushing |
| playing shofar           | dog growling  |
| airplane flyby            | splashing water |
| playing trumpet           | wind chime    |
| wind chime                | striking bowling |
| Others                   | Others        |
| cat purring              | dog barking   |
| dog barking              | race car      |
| race car                 | singing bowl  |
| singing bowl             | vacuum cleaner cleaning floors |
| vacuum cleaner cleaning floors | toilet flushing |
| toilet flushing          | dog growling  |
| dog growling             | splashing water |
Class-wise performance on VGG-Sound

| Slow stream | Fast stream |
|-------------|-------------|
| Animals     | Percussive sounds |
|             | sewing choir   |
|             | voices         |
|             | Others         |

- Animals: baltimore oriole calling, cheetah chirrup, zebra braying, dinosaurs bellowing, horse neighing, black capped chickadee calling, cat hissing, cuckoo bird calling, mosquito buzzing, bull bellowing, whale calling

- Percussive sounds: footsteps on snow, snake rattling, tap dancing, car engine knocking, woodpecker pecking tree, chopping wood, people clapping, lawn mowing, typing on typewriter, opening or closing car doors, playing tennis, railroad car, playing tympani, playing drum kit, playing vibraphone, popping pop corn

- Voices: singing choir, people cheering, people crowd, child speech, baby laughter

- Others: cat purring, dog barking, race car, singing bowl, vacuum cleaner cleaning floors, toilet flushing, dog growling, splashing water
Qualitative Results - EPIC-KITCHENS

GT: wash countertop
Slow: wash countertop
Fast: wash countertop
Slow-Fast: wash countertop

Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).
Qualitative Results - EPIC-KITCHENS

GT: squeeze orange
Slow: press orange
Fast: wash plate
Slow-Fast: squeeze orange
Qualitative Results - EPIC-KITCHENS

GT: cut tomato
Slow: cut tomato
Fast: cut carrot
Slow-Fast: cut pepper
Qualitative Results - EPIC-KITCHENS

GT: put package
Slow: put cheese
Fast: put package
Slow-Fast: put biscuit
Qualitative results - VGG-Sound

|         | people clapping |
|---------|-----------------|
| GT:     | people clapping |
| Slow:   | people clapping |
| Fast:   | people clapping |
| Slow-Fast: | people clapping |
Qualitative results - VGG-Sound

GT: people sneezing
Slow: cat purring
Fast: people coughing
Slow-Fast: people sneezing
Qualitative results - VGG-Sound

GT:  sliding door
Slow: sliding door
Fast: typing on typewriter
Slow-Fast: typing on typewriter
Qualitative results - VGG-Sound

GT: chopping wood
Slow: hammering nails
Fast: chopping wood
Slow-Fast: hammering nails
Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

**Links**

- Project webpage: [https://ekazakos.github.io/auditoryslowfast/](https://ekazakos.github.io/auditoryslowfast/)

- Code & models: [https://github.com/ekazakos/auditory-slow-fast](https://github.com/ekazakos/auditory-slow-fast)