University of BRISTOL

UNIVERSITY OF OXFORD

SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, Dima Damen

ICASSP2021
TORONTO
Canada
June 6-11, 2021
Metro Toronto Convention Centre

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

Audio Signal – EPIC-KITCHENS

- Hand-object interactions

“crush bag”

Damen et al. (2020). Rescaling Egocentric Vision, arXiv

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

Audio Signal – EPIC-KITCHENS

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

“turn-on blender”

Damen et al. (2020). Rescaling Egocentric Vision, arXiv

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

Audio Signal – EPIC-KITCHENS

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

“rinse bell pepper”

Damen et al. (2020). Rescaling Egocentric Vision, arXiv

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

Audio Signal – EPIC-KITCHENS

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

“chop garlic cloves”

Damen et al. (2020). Rescaling Egocentric Vision, arXiv

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

Audio Signal – VGG-Sound

- Harmonic sounds

“thunder”

Chen et al. (2020). VGGSound: A Large-scale Audio-Visual Dataset. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

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

Audio Signal – VGG-Sound

- Harmonic sounds

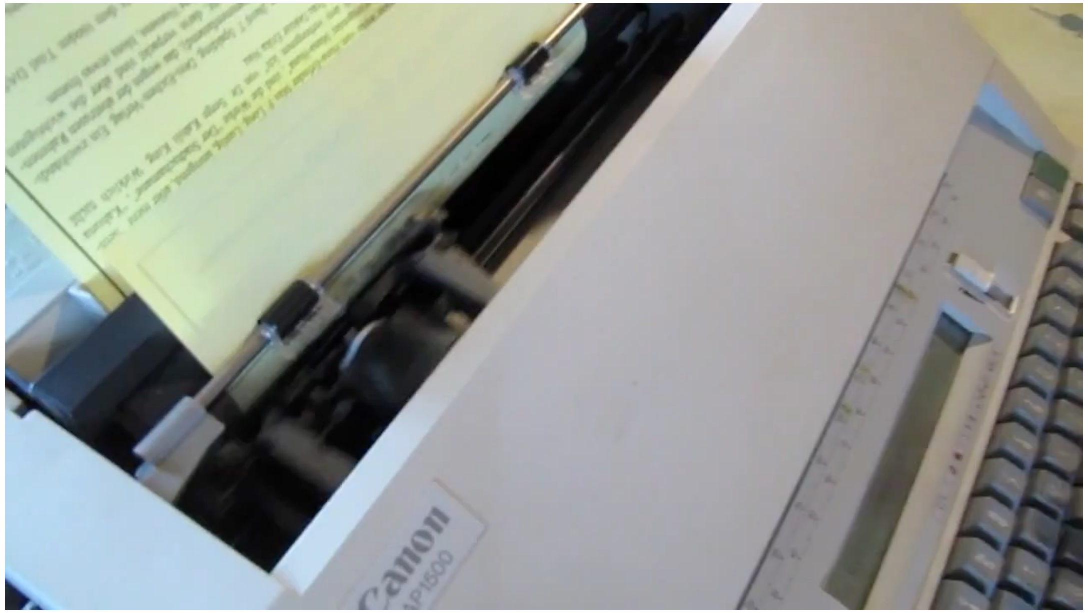
“canary calling”

Chen et al. (2020). VGGSound: A Large-scale Audio-Visual Dataset. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

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

Audio Signal – VGG-Sound

- Harmonic sounds
- Percussive sounds



“typing on typewriter”

Chen et al. (2020). VGGSound: A Large-scale Audio-Visual Dataset. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

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

Audio Signal – VGG-Sound

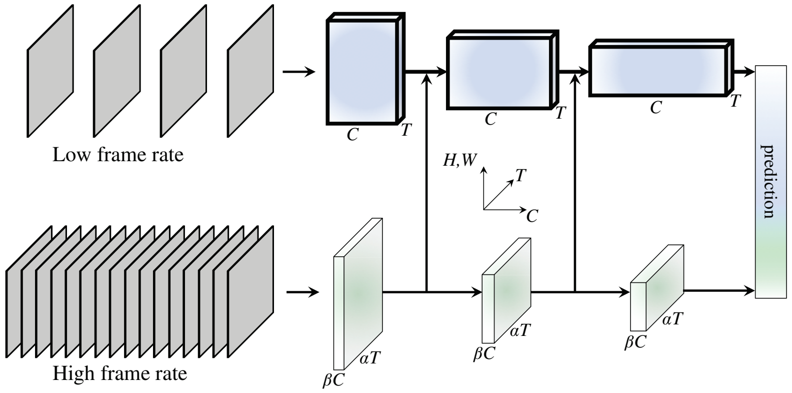
- Harmonic sounds
- Percussive sounds

“playing tennis”

Chen et al. (2020). VGGSound: A Large-scale Audio-Visual Dataset. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

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]

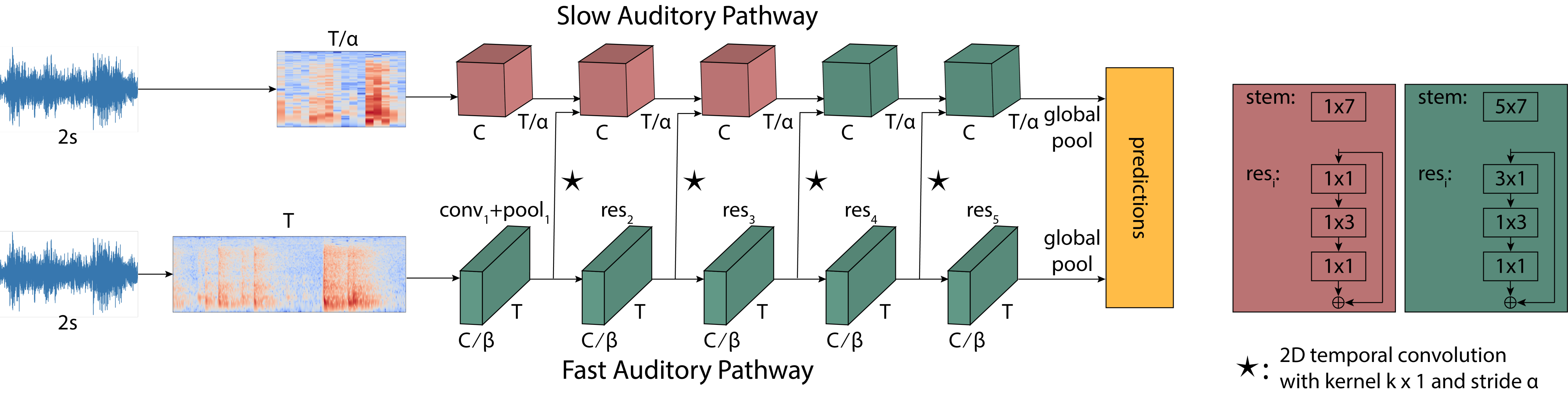


[1] Santoro et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLOS Computational Biology

[2] Feichtenhofer et al. (2019). SlowFast Networks for Video Recognition. International Conference on Computer Vision (ICCV)

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

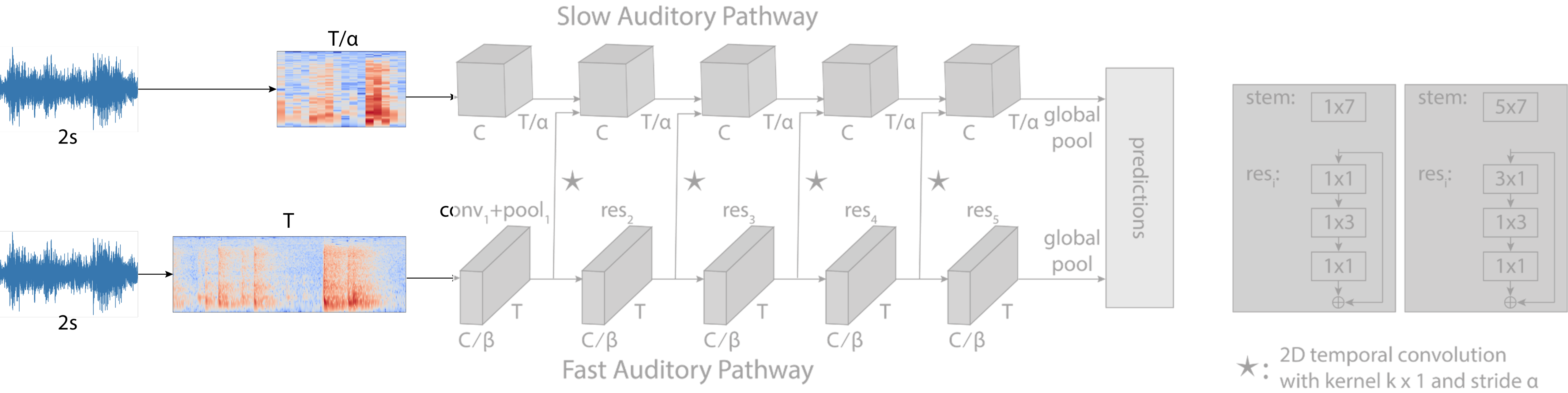
Auditory Slow-Fast Network



★: 2D temporal convolution with kernel k x 1 and stride α

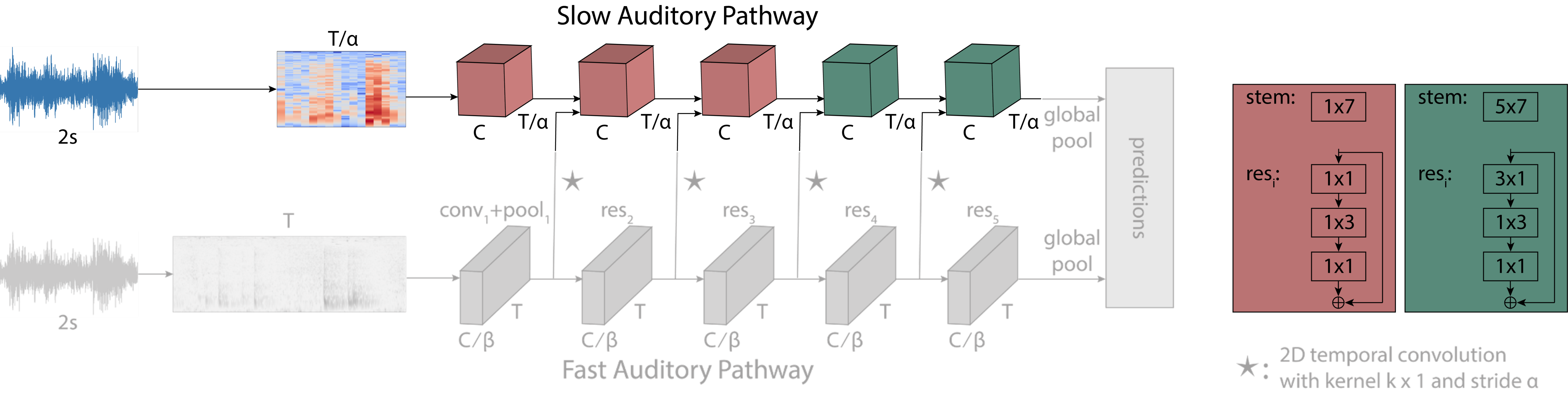
Kazakos et al. (2021). SLOW-FAST AUDITORY STREAMS FOR AUDIO RECOGNITION. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Auditory Slow-Fast Network



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

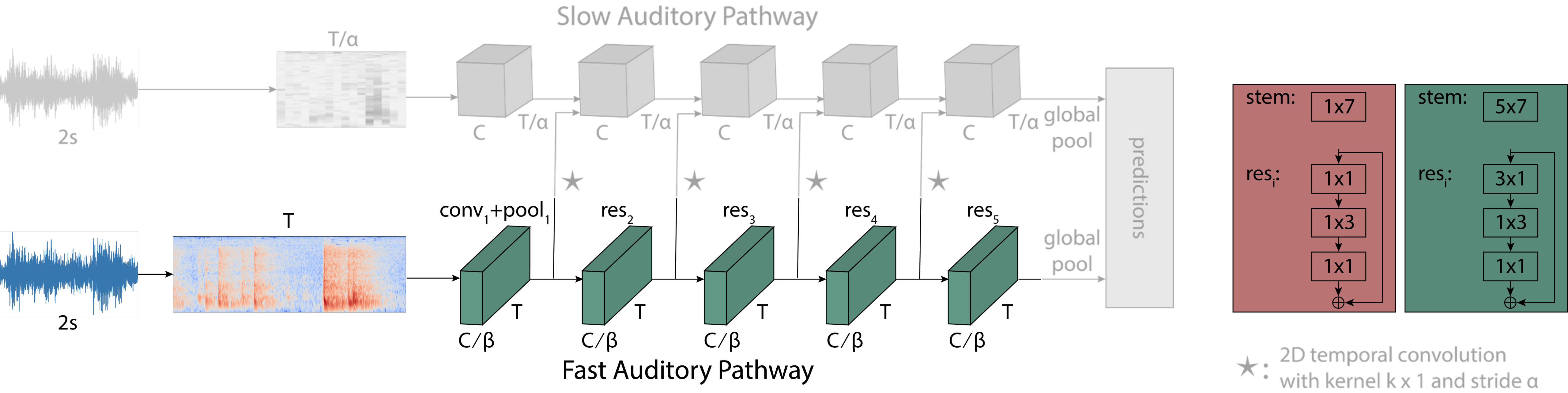
Auditory Slow-Fast Network



- Slow has low temporal precision and large amount of channels

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

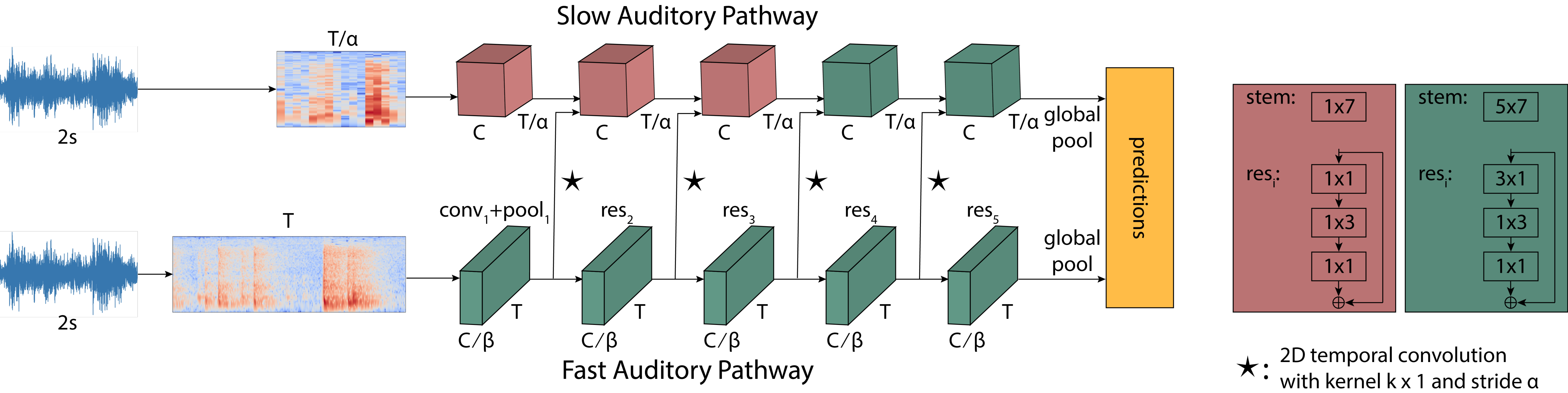
Auditory Slow-Fast Network



- 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).

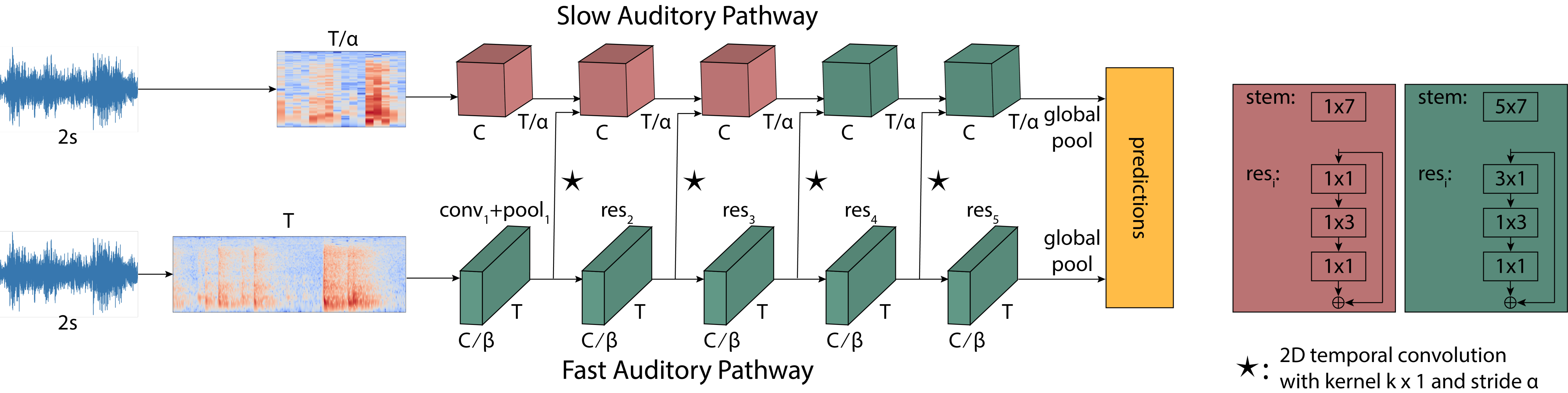
Auditory Slow-Fast Network



- 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).

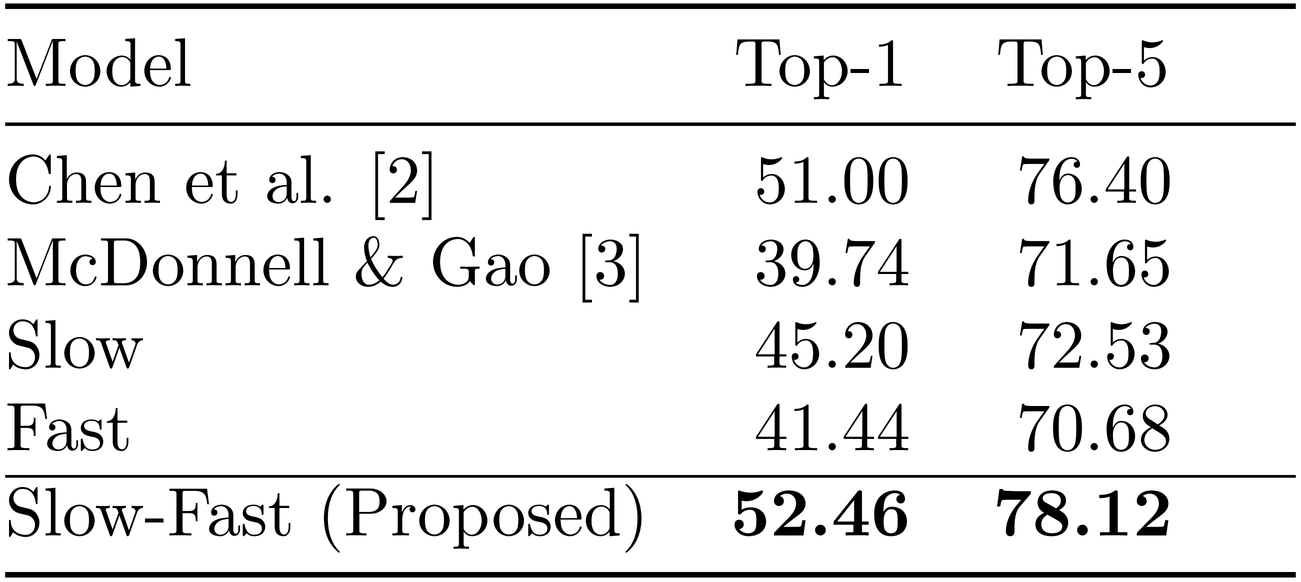
Auditory Slow-Fast Network



★: 2D temporal convolution with kernel k x 1 and stride α

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

Results: VGG-Sound

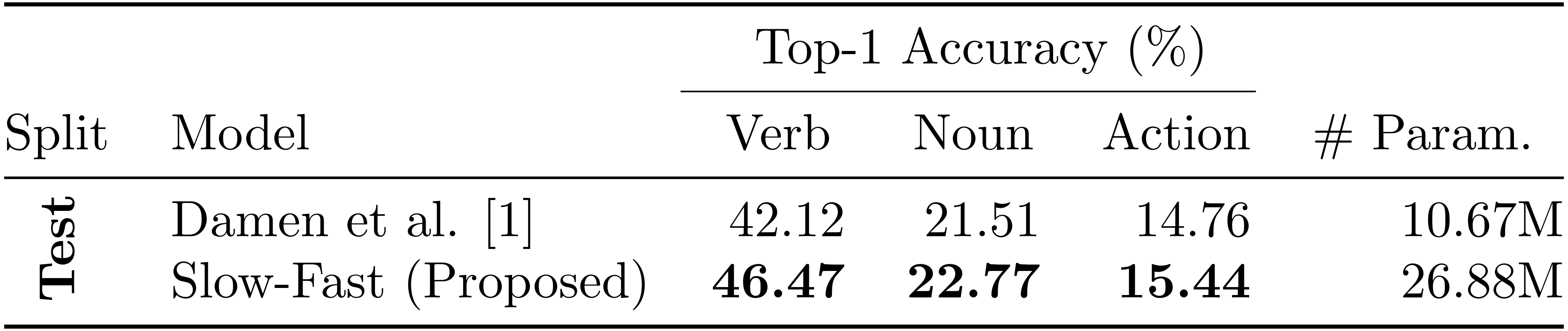


Model	Top-1	Top-5
Chen et al. [2]	51.00	76.40
McDonnell & Gao [3]	39.74	71.65
Slow	45.20	72.53
Fast	41.44	70.68
Slow-Fast (Proposed)	**52.46**	**78.12**

[2] Chen et al. (2020). VGGSound: A Large-scale Audio-Visual Dataset. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

[3] M. McDonnell and W. Gao. (2020). Acoustic scene classification using deep residual networks with late fusion of separated high and low frequency paths. (ICASSP)

Results: EPIC-KITCHENS

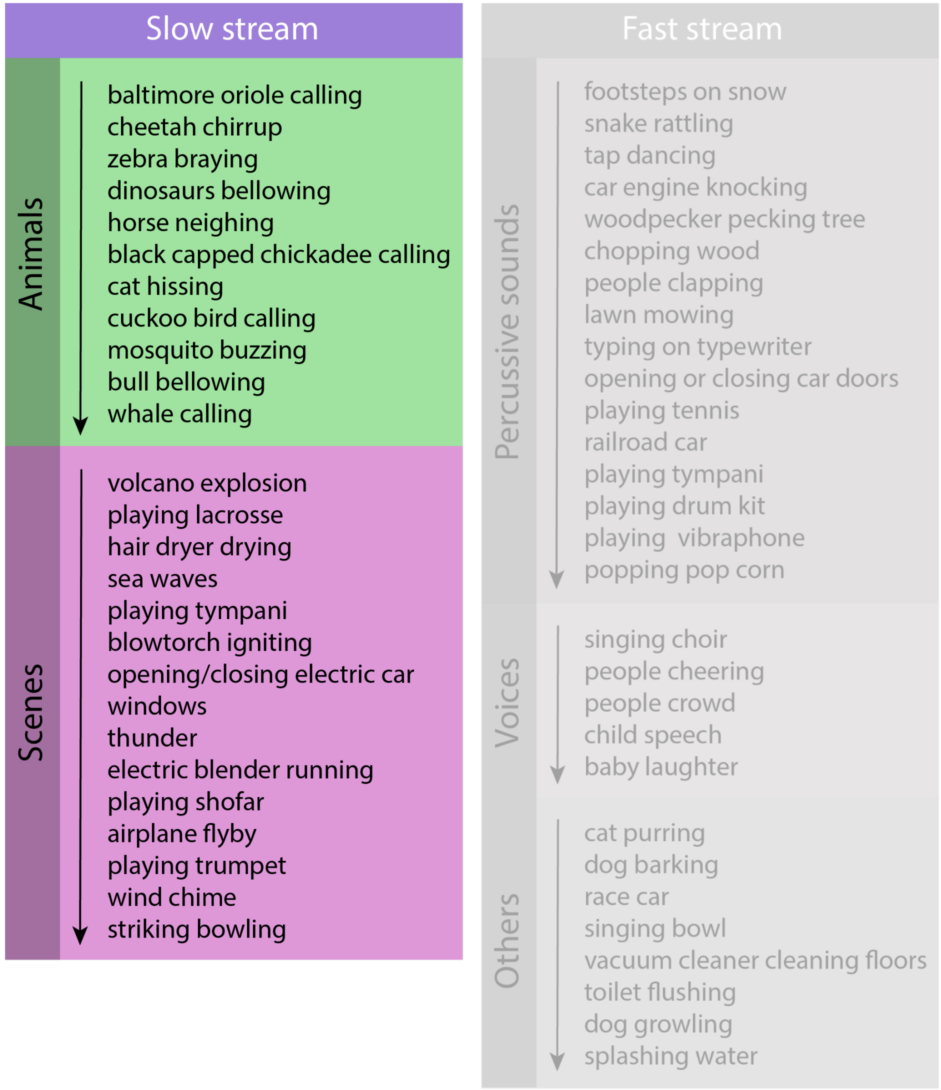


		Top-1 Accuracy (%)			
Split	Model	Verb	Noun	Action	# Param.
Test	Damen et al. [1]	42.12	21.51	14.76	10.67M
	Slow-Fast (Proposed)	**46.47**	**22.77**	**15.44**	26.88M

[1] Damen et al. (2020). Rescaling Egocentric Vision, arXiv

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

Class-wise performance on VGG-Sound



Slow stream

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

Scenes

volcano explosion
playing lacrosse
hair dryer drying
sea waves
playing tympani
blowtorch igniting
opening/closing electric car
windows
thunder
electric blender running
playing shofar
airplane flyby
playing trumpet
wind chime
striking bowling

Fast stream

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

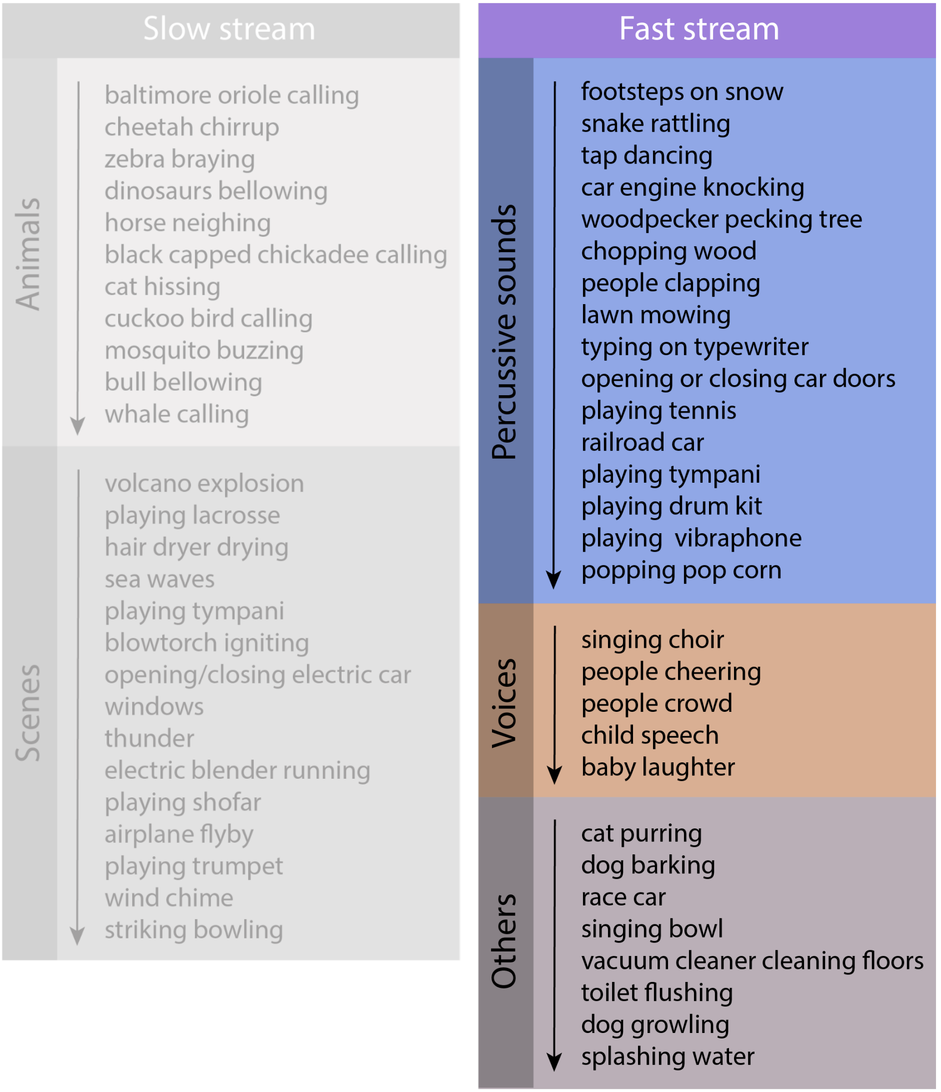
singing choir
people cheering
people crowd
child speech
baby laughter

Others

cat purring
dog barking
race car
singing bowl
vacuum cleaner cleaning floors
toilet flushing
dog growling
splashing water

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

Class-wise performance on VGG-Sound



Slow stream

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

Scenes

- volcano explosion
- playing lacrosse
- hair dryer drying
- sea waves
- playing tympani
- blowtorch igniting
- opening/closing electric car windows
- thunder
- electric blender running
- playing shofar
- airplane flyby
- playing trumpet
- wind chime
- striking bowling

Fast stream

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

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:	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

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:	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

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

Qualitative results - VGG-Sound

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

Links

- Project webpage: https://ekazakos.github.io/auditoryslowfast/
- Code & models: https://github.com/ekazakos/auditory-slow-fast