All-in-One Transformer: Unifying Speech Recognition, Audio Tagging, and Event Detection

Niko Moritz, Gordon Wichern, Takaaki Hori, Jonathan Le Roux

Interspeech 2020

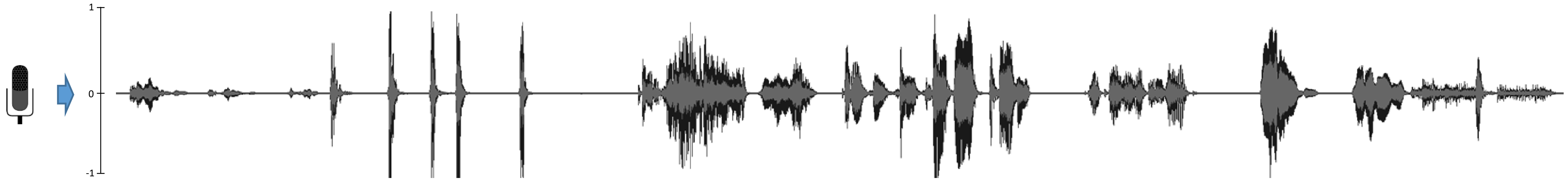
MITSUBISHI ELECTRIC RESEARCH LABORATORIES (MERL)
Cambridge, Massachusetts, USA
http://www.merl.com

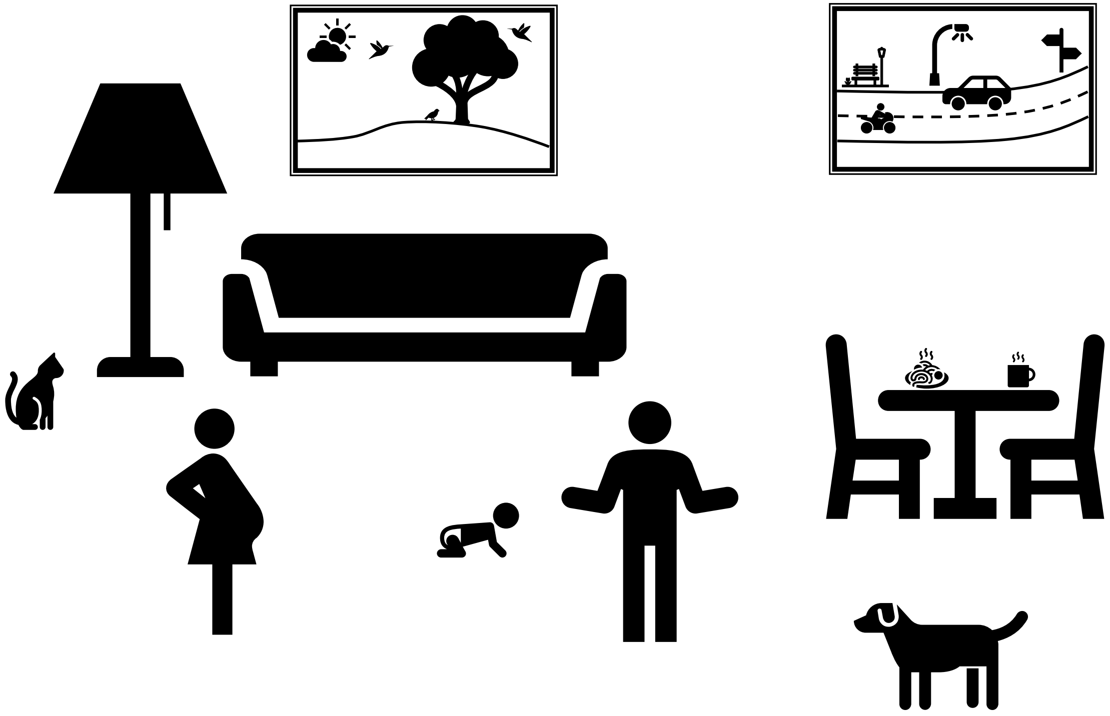
All-in-One Transformer: Unifying ASR, AT, and AED

- Motivation:
  - Automatic speech recognition (ASR), audio event detection (AED), and audio tagging (AT) are traditionally treated as separate problems with custom-made solutions.
  - In contrast, the human auditory system uses a single (binaural) pathway to process sound signals from different sources.

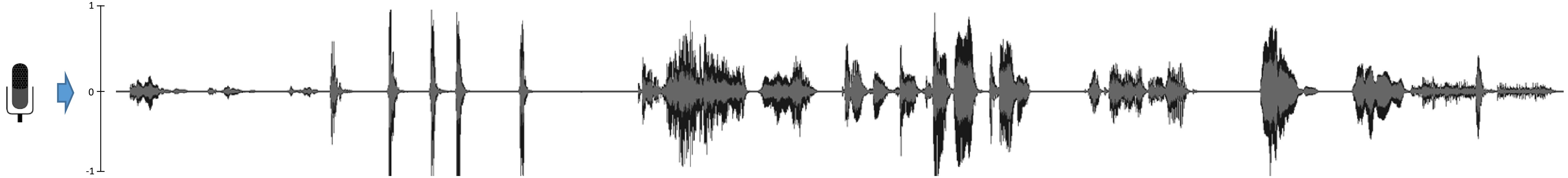
- Investigated Questions:
  - Can we develop a system that moves closer to the versatility of the human auditory system?
  - Can training on multiple heterogeneous tasks lead to a single system with performance similar to or better than systems developed independently for each task?
  - Can a single system successfully handle multiple tasks with widely varying characteristics, large length discrepancies, and w/ or w/o monotonicity?

Acoustic Scene





Audio Tagging (AT)



cat meowing

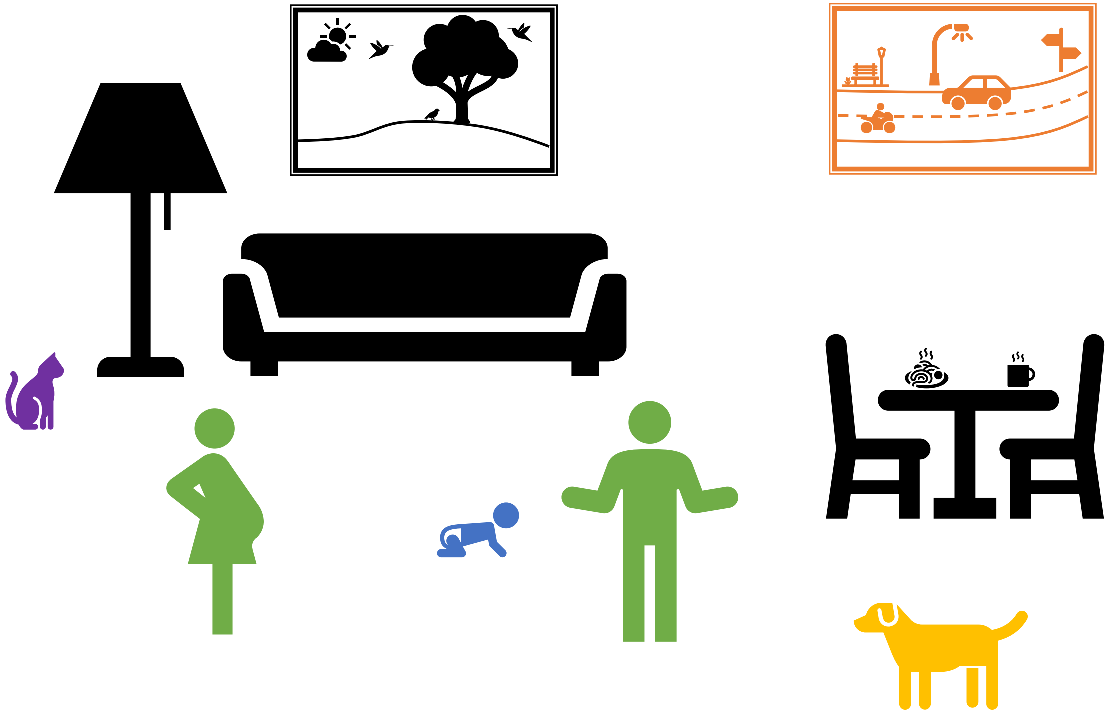
speech

baby crying

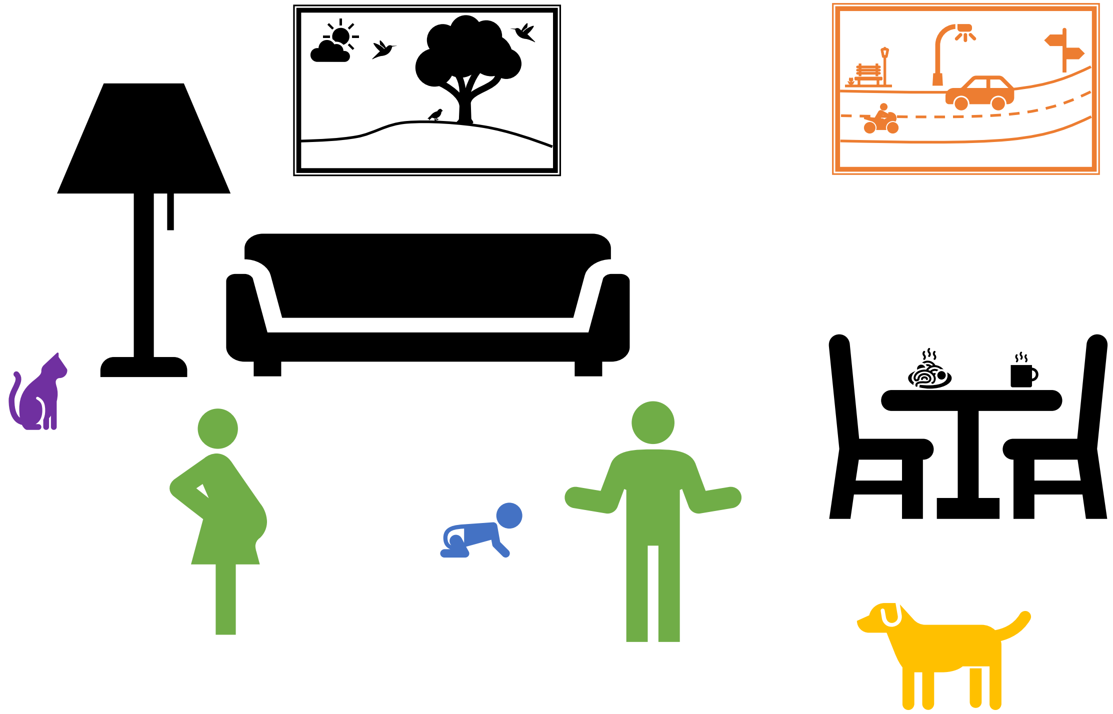
dog barking

traffic sounds

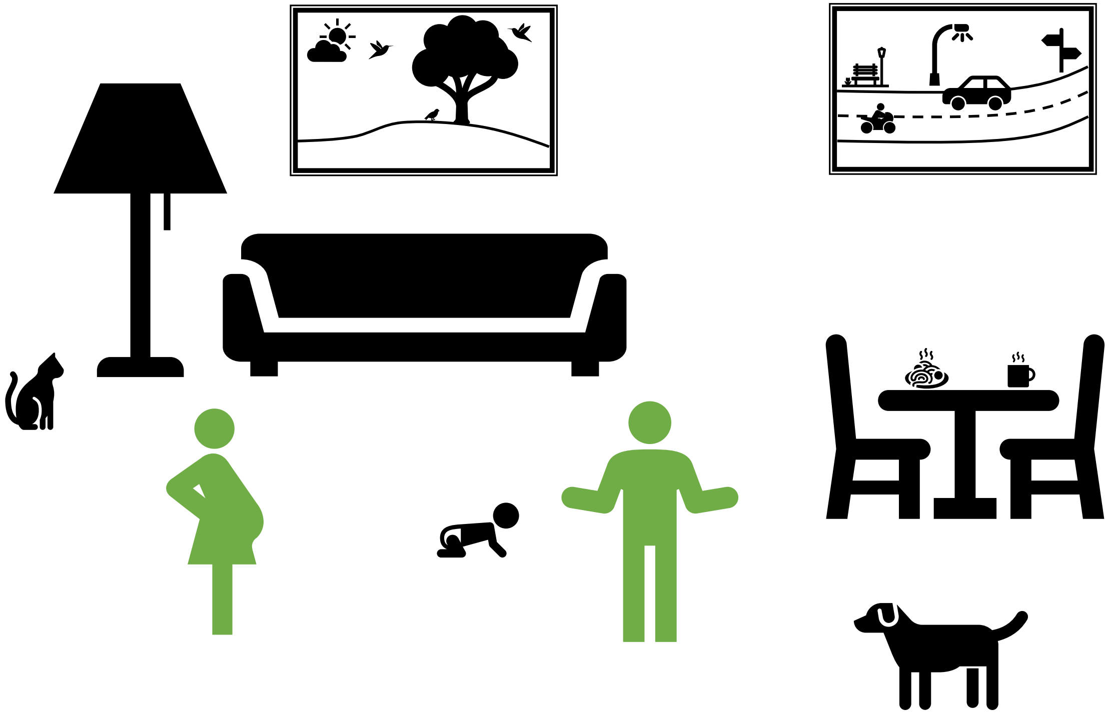
home environment



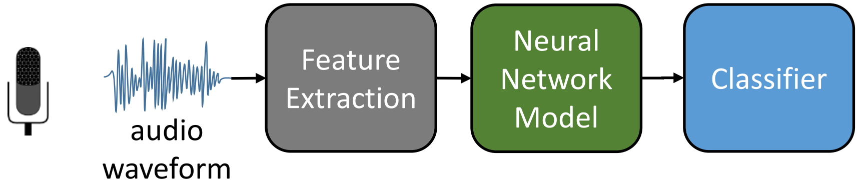
Acoustic Event Detection (AED)



Automatic Speech Recognition (ASR)



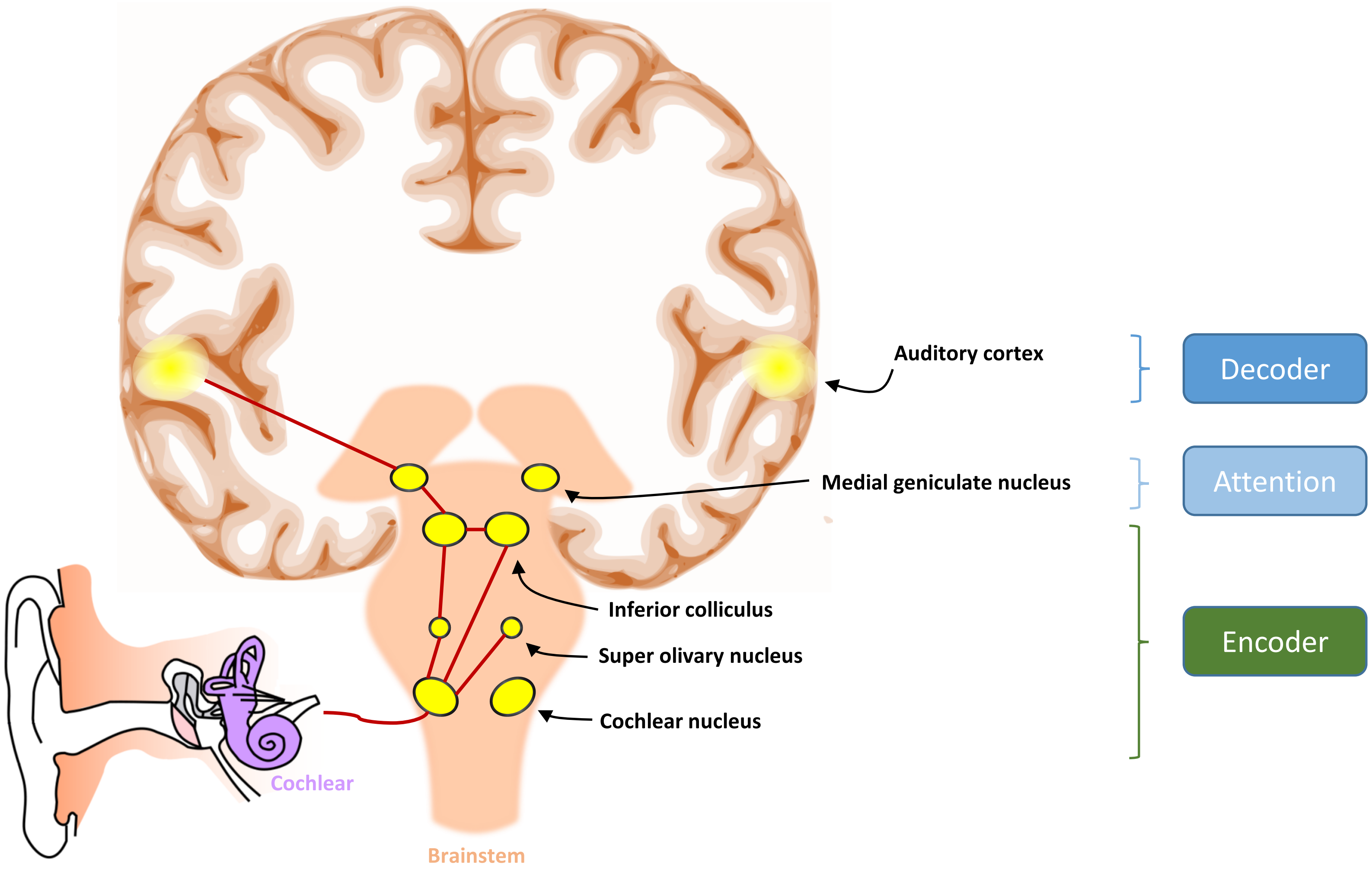
Baseline System Architectures of DCASE 2019 Task 1, 2, 4, and 5



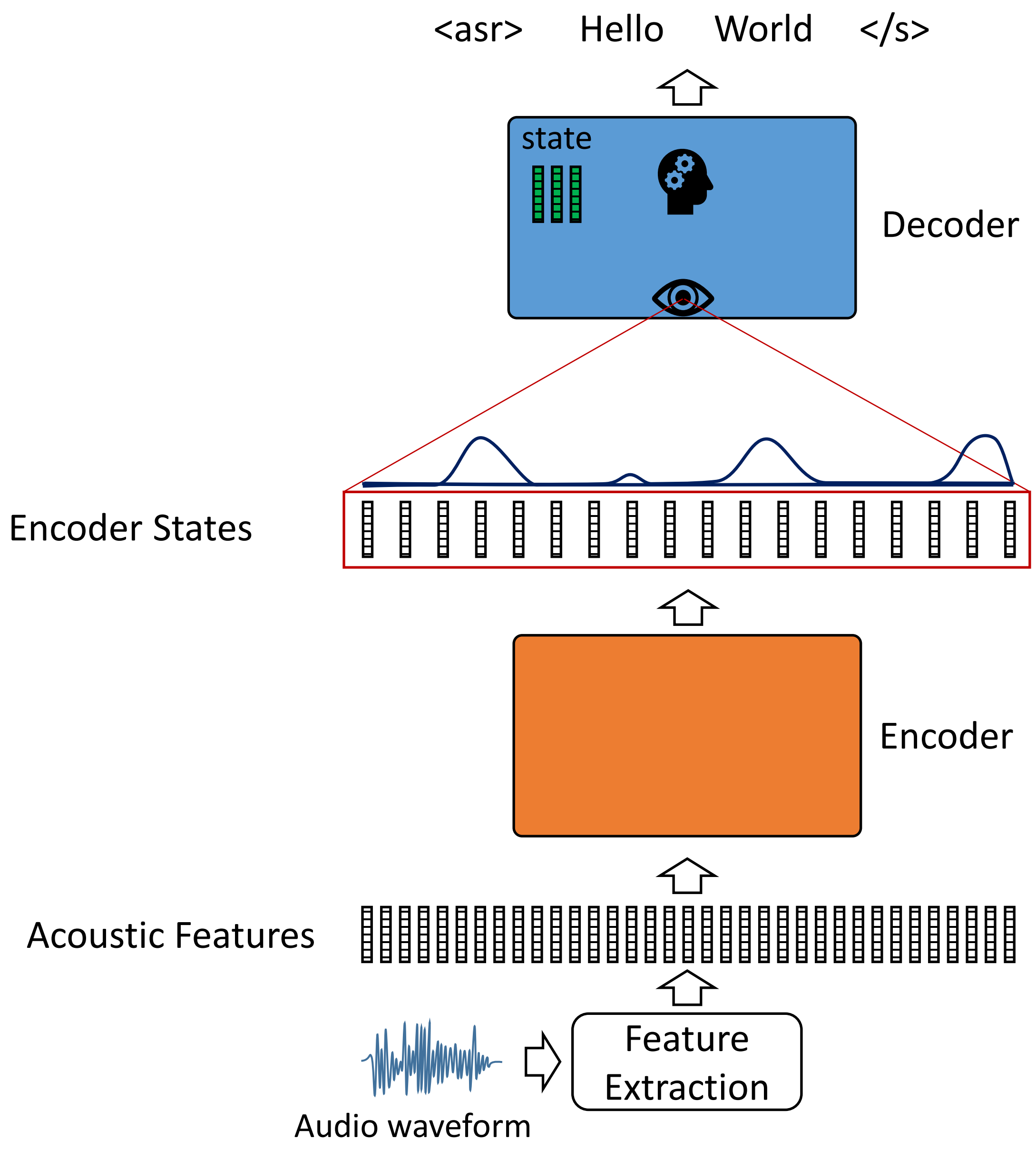
Audio sampling rates	Feature Extraction: Log-Mel Spectral Engergies	Neural Network Models	Classifier Methods
16 kHz, 22.05 kHz 44.1 kHz, or 48 kHz;	40, 64, 96, or 128-dimensional; Different window and hop sizes;	CNNs, DNNs, RNNs, ...; MobileNet v1; VGGish;	Logistic regression; Max and average pooling; Attention-based pooling Clip- and frame-leve classification;

- A. Mesaros, T. Heittola, and T. Virtanen, “A multi-device dataset for urban acoustic scene classification,” in Proc. of the DCASE Workshop, 2018.
- E. Fonseca, M. Plakal, F. Font, D. Ellis, and X. Serra, “Audio tagging with noisy labels and minimal supervision,” in Proc. of the DCASE Workshop, 2019.
- L. JiaKai, “Mean teacher convolution system for DCASE 2018 task 4,” in Proc. of the DCASE Workshop, 2018.
- J. Bello, C. Silva, O. Nov, R. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, “SONYC: a system for monitoring, analyzing, and mitigating urban noise pollution,” Communications of the ACM, 2019.
- S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, “CNN architectures for large-scale audio classification,” in Proc. IEEE ICASSP, 2017.

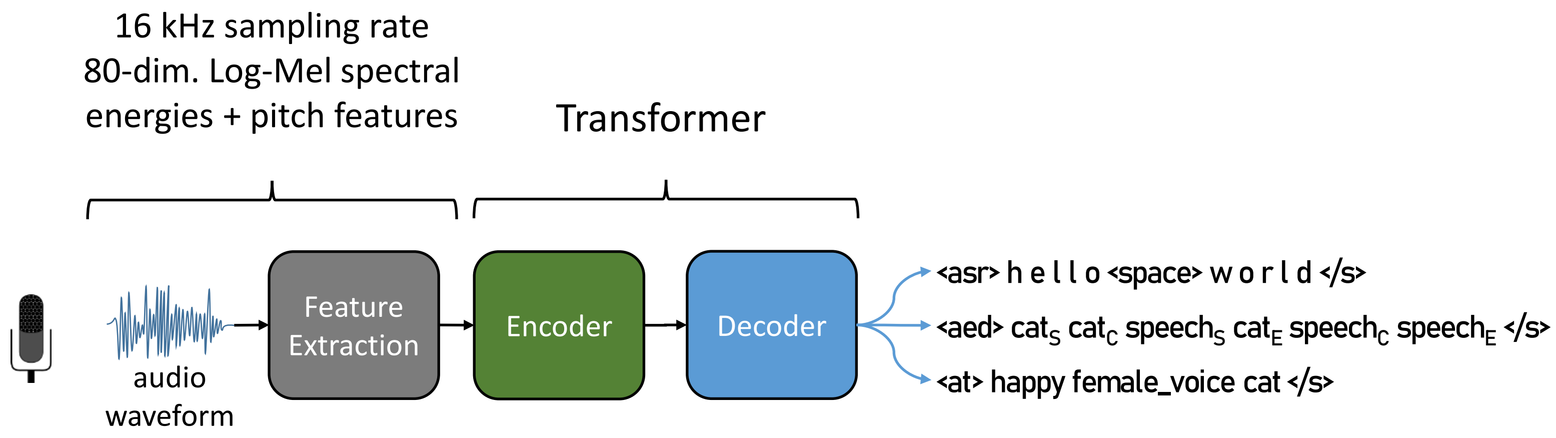
The Auditory Pathway



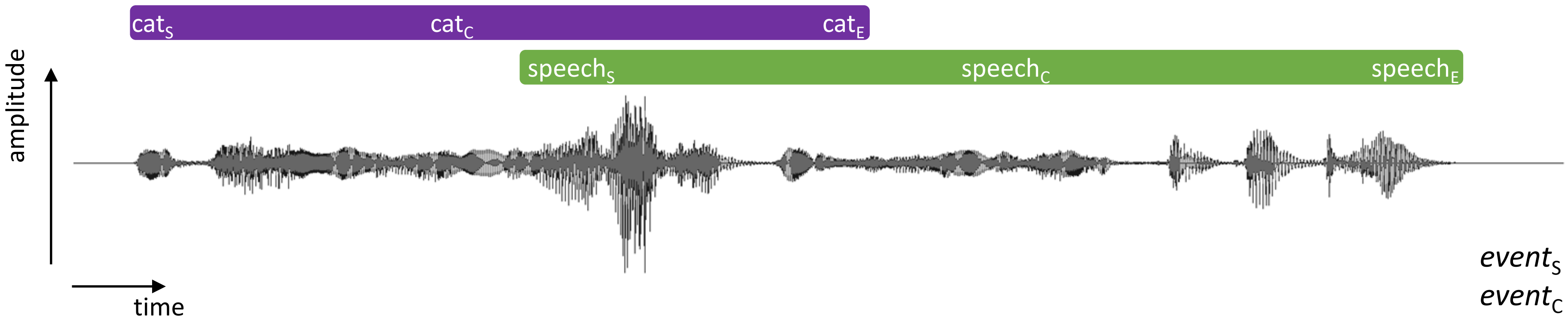
Attention-based Encoder-Decoder



System Architecture: Attention-Based Encoder-Decoder



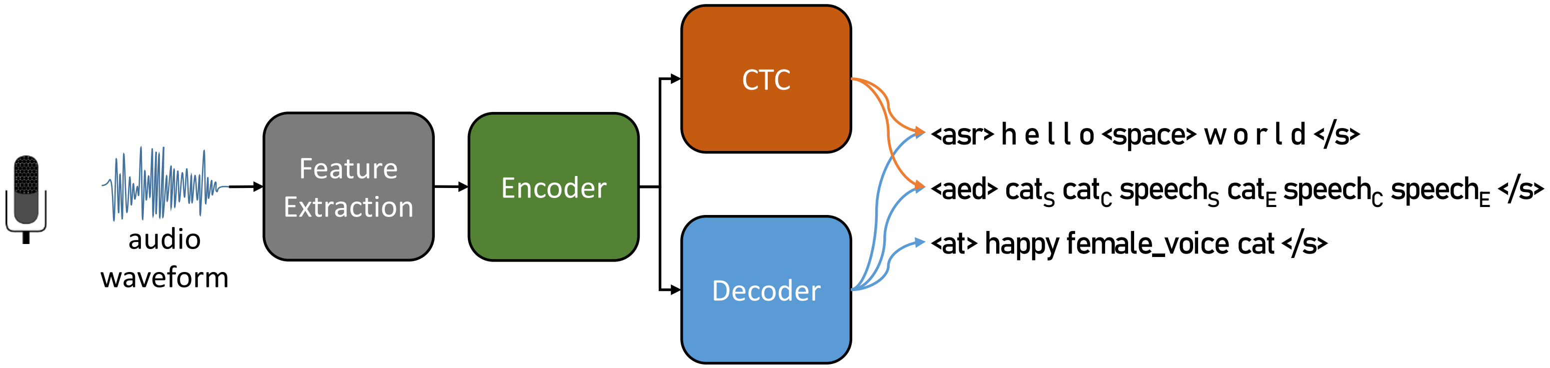
Acoustic Event Detection



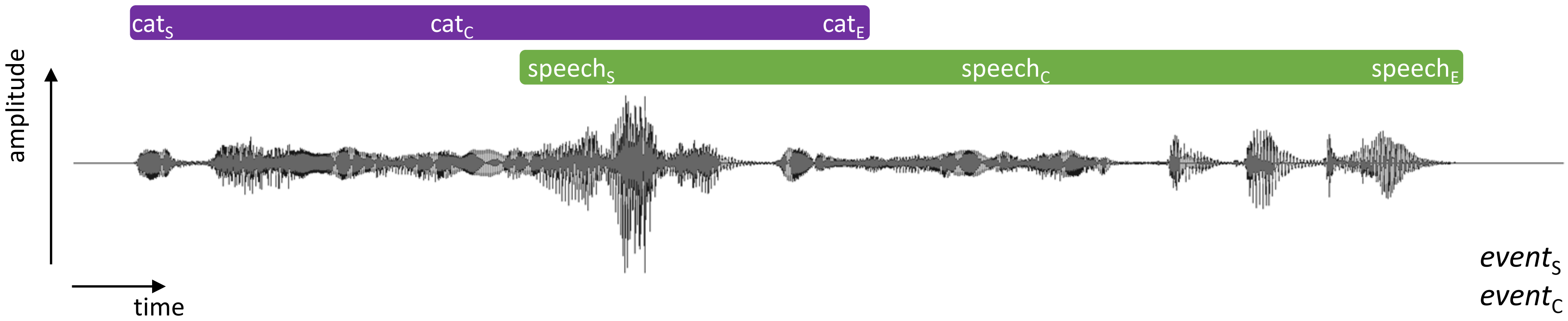
$event_S$ => event start
$event_C$ => event continues
$event_E$ => event end

System Architecture: Hybrid CTC-Transformer

Connectionist Temporal Classification (CTC)



Acoustic Event Detection



$event_S$ => event start
$event_C$ => event continues
$event_E$ => event end

Data Sets

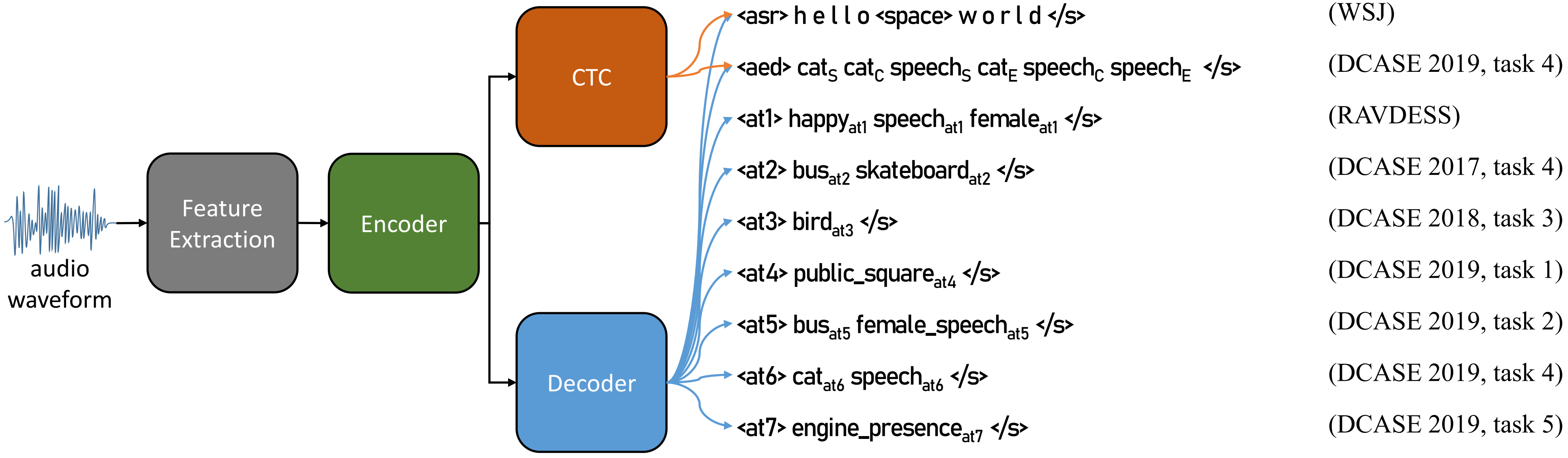
- Automatic Speech Recognition (ASR)
  - Wall Street Journal (**WSJ**): "**Read English newspapers**".
    Train: 81h; Dev.: 1.1h; Test: 0.7h

- Acoustic Event Detection (AED)
  - DCASE 2019 task 4 (**DCASE19-4**): "**Sound event detection in domestic environments**".
    Train: 5.7h; Dev.: 2.9h; Test: 1.9h

Data Sets

- Audio Tagging (AT):
  - DCASE 2017 task 4 (**DCASE17-4**): "**Large-scale weakly supervised sound event detection for smart cars**".
    Train: 140h; Dev.: 1.3h; Test: 3h
  - DCASE 2018 task 3 (**DCASE18-3**): "**Bird audio detection**".
    Train: 99h; Dev.: n/a; Test: n/a
  - DCASE 2019 task 1 (**DCASE19-1**): "**Audio scene classification**".
    Train: 25.5h; Dev.: 11.6h; Test: 9.8h
  - DCASE 2019 task 2 (**DCASE19-2**): "**Audio tagging with noisy labels and minimal supervision**".
    Train: 90.8; Dev.: 3.1h; Test: 9.8h
  - DCASE 2019 task 4 (**DCASE19-4**): "**Sound event detection in domestic environments**".
    Train: 9.8h; Dev.: 2.9h; Test: 1.9h
  - DCASE 2019 task 5 (**DCASE19-5**): "**Urban sound tagging**".
    Train: 4.4h; Dev.: 1.2h; Test: 0.7h
  - The Ryerson Audio-Visual Database of Emotional Speech and Song (**RAVDESS**): Recognition of "**emotion**" + "**vocal channel**" + "**gender**"
    Train: 2.8h; Dev.: n/a; Test: n/a

All-in-One (AIO) Transformer



Audio Tagging – Results

Micro-averaged F1-scores [%]

				DCASE19							DCASE18		DCASE17		RAVDESS	
	Training data			Task 1	Task 2		Task 4		Task 5		Task 3		Task 4			
System	AT	AED	ASR	dev	dev	test	dev	test	dev	test	dev	test	dev	test	dev	test
Baseline Systems	single	✗	✗	62.5	39.8	38.8	71.4	66.8	73.0	68.9	n/a	n/a	19.0	29.3	n/a	n/a

Automatic Speech Recognition – Results

	Training data			Word Error Rates [%] WSJ	
System	AT	AED	ASR	dev	test
CTC-Transformer	✗	✗	✓	7.7	5.0
CTC-Transformer	✗	✓	✓	7.8	5.0
AIO Transformer	multi	✓	✓	7.5	5.1

Multi-condition training using DEMAND and NOISEX data sets.
Noisy test conditions using the DCASE data sets.

* Multi-condition training

Acoustic Event Detection – Results

				F1-scores [%]			
	Training data			Event-based		Segment-based	
System	AT	AED	ASR	dev	test	dev	test
Baseline system	✗	✓	✗	29.0	24.0	58.5	54.8
CTC-Transformer	✗	✓	✗	16.0	10.6	43.8	34.8

Event-based F1-scores: 200 ms collar for on- and offsets
Segement-based F1-scores: 1 sec. long segments

* Multi-condition training

Conclusions

- ASR, AED, and AT tasks can be unified under a single system architecture, where model parameters are shared across all tasks.
- Multi-task learning has shown to improve results for individual tasks.
- The AIO Transformer model has achieve competitive or better results compared to all tested DCASE challenge baseline systems, as well as to an ASR baseline system of similar architecture.
- The proposed system can be used to perform the *total transcription* of an acoustic scene, i.e., a single system can be used to transcribe speech as well as other acoustic events.

MITSUBISHI
ELECTRIC
Changes for the Better