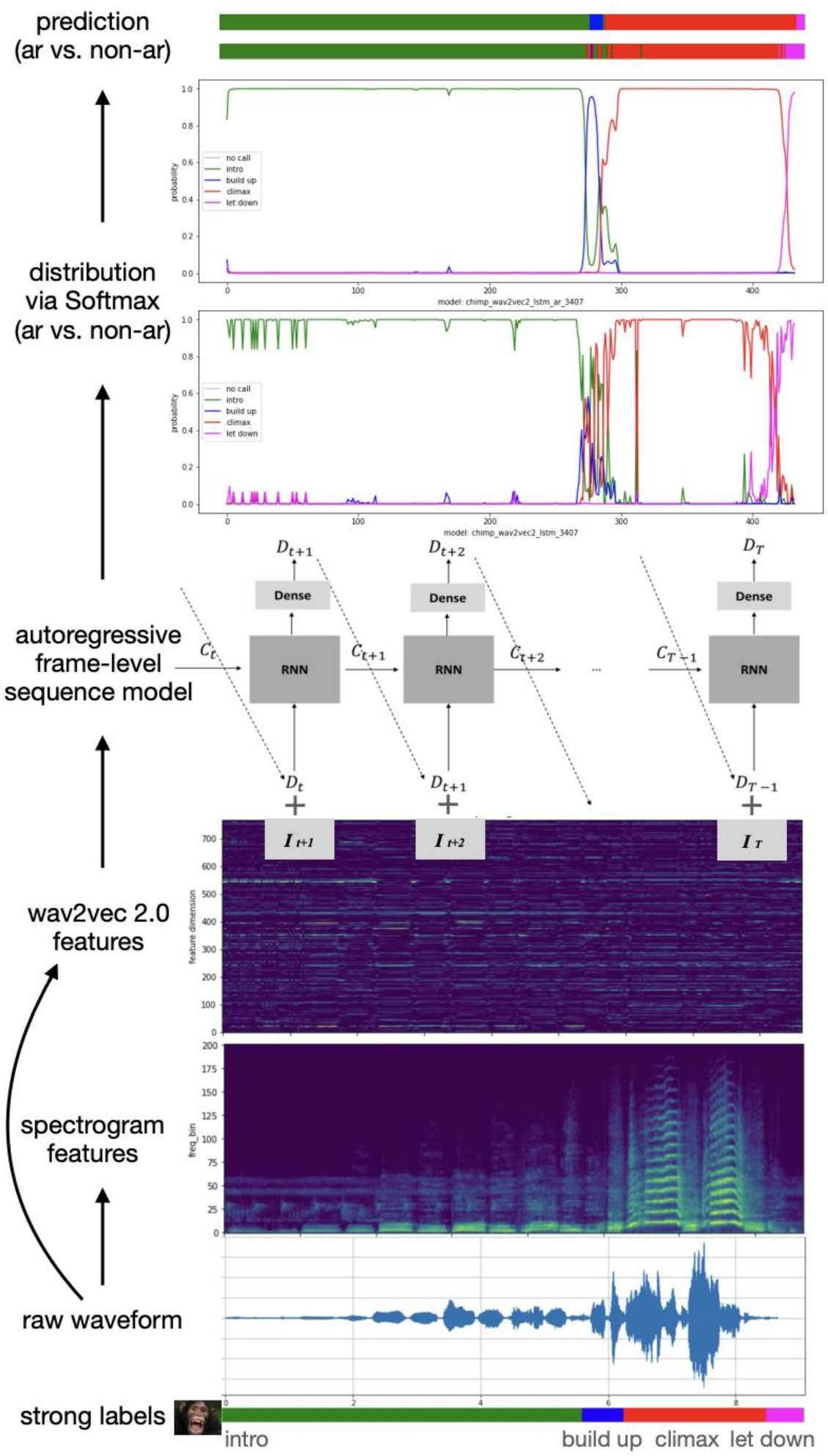
Automatic Sound Event Detection and Classification of Great Ape Calls Using Neural Networks

Zifan Jiang[1,2], Adrian Soldati[1,3], Isaac Schamberg[2], Adriano R. Lameira[4], Steven Moran[1,5]
[1] University of Neuchâtel, [2] University of Zurich, [3] University of St Andrews,
[4] University of Warwick, [5] University of Miami
Contact Email: j22melody@gmail.com

TL;DR

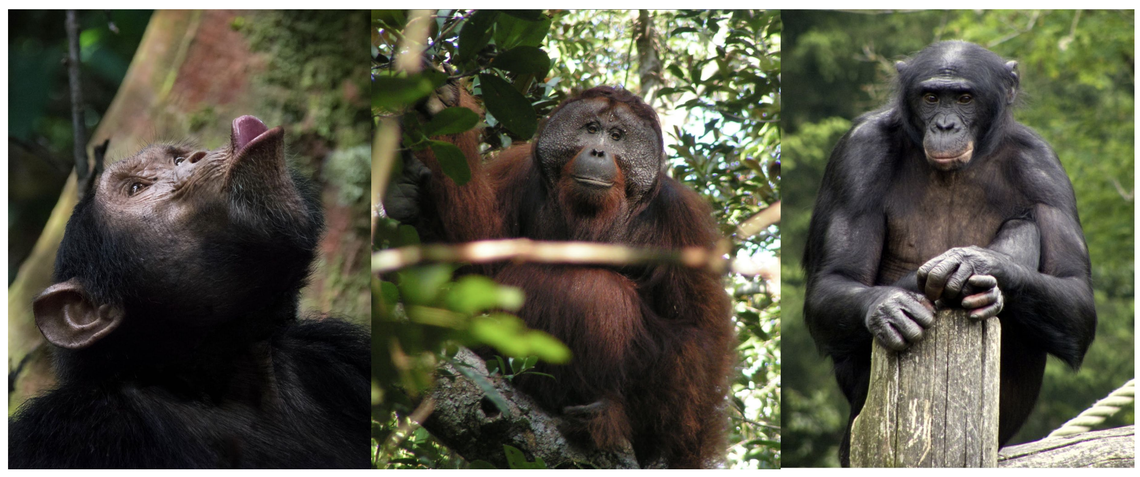
- Human annotation is expensive - **automatically detect and classify great ape calls** from continuous audio recordings.
- **Three data sets** of different great ape lineages collected during field research: chimpanzees, orangutans, and bonobos.
- **Wav2vec 2.0**, pretrained on 1000 hours human speech, **transfers** surprisingly well as an acoustic feature extractor used with LSTM.

Method



Data

data set	# audio clips	mean / call / total duration
chimpanzee	235	~ 8s / 1,955s / 1,964s
orangutan	65	~ 74s / 2,793s / 4,817s
bonobo	28	~ 24s / 62s / 677s



Results

ID	data	feature	model	test acc.	test f1
Explore the best feature and model combination					
E1	chimp	waveform	lstm (baseline)	51.0 ± 3.6	34.7 ± 4.3
E1.1	chimp	spectrogram	lstm	58.7 ± 4.7	53.9 ± 5.4
E2	chimp	wav2vec2	lstm	79.3 ± 2.3	77.9 ± 3.6
E2.1	chimp	wav2vec2	transformer	75.3 ± 0.6	72.1 ± 0.5
Explore the hyper-parameters					
E3.1	chimp	wav2vec2	lstm (E2 + *batch_size* = 4)	67.7 ± 4.0	69.6 ± 4.0
E3.2	chimp	wav2vec2	lstm (E2 + *batch_size* = 8)	62.0 ± 4.4	61.5 ± 4.0
E3.3	chimp	wav2vec2	lstm (E2 + *dropout* = 0.2)	78.0 ± 1.7	76.8 ± 2.7
E3.4	chimp	wav2vec2	lstm (E2 + *dropout* = 0.1)	78.7 ± 2.9	77.3 ± 3.9
E3.5	chimp	wav2vec2	lstm (E2 - *balance_weights*)	79.3 ± 2.3	78.3 ± 3.6
Explore autoregressive modeling					
E4	chimp	wav2vec2	lstm (E2 + autoregressive)	85.7 ± 2.1	85.6 ± 2.5
Extend to orangutan long calls and a binary setting					
E5	orang	wav2vec2	lstm (= E4)	81.7 ± 3.1	82.0 ± 2.6
E5.1	orang	wav2vec2	lstm (E5 + binary target)	92.0 ± 1.0	91.9 ± 1.1
Extend to bonobo calls and a binary setting					
E6	bonobo	wav2vec2	lstm (= E4)	83.7 ± 3.8	82.3 ± 2.2
E6.1	bonobo	wav2vec2	lstm (E6 + binary target)	87.7 ± 3.5	87.8 ± 2.9
Zero-shot transferring from orangutan to bonobo					
E7	bonobo	wav2vec2	lstm (= E5.1)	72.0 ± 4.0	74.2 ± 3.1

Table 2: We run all experiments three times based on different random seeds and report the mean and standard deviation. acc. stands for frame-level accuracy, f1 stands for the frame-level average F1- score weighted by the number of true instances per class. For hyper-parameters, we start E1 with batch_size = 1, dropout = 0.4 and keep them by default, if not otherwise specified in the table.

Discussion

- Wav2vec 2.0 transfers from human speech (high resource) to great ape calls (low resource) - what about **other animals**?
- Do the same observations hold if we move from the vocal-auditory channel to the **manual-visual channel** (gestures)?
- A broader picture of **decoding** the communication systems of non-human animals - e.g., the Earth Species Project.

We gratefully acknowledge funding from: UK Research & Innovation, Future Leaders Fellowship (MR/T04229X/1; ARL), Swiss National Science Foundation (PCEFP1_186841: SM, ZJ; 310030_185324: AS), St Leonard College (AS), and Swissuniversities (AS). We thank also Klaus Zuberbühler, Josep Call, and the field assistants of the Budongo Conservation Field Station.