Systems description

- *Number of parameters in the model:* 8,000,000 (~8 M)
- *FLOPS:* 11.45 GFlops
- *Number of training steps:*8632 steps per epoch.
- *Latency (with hardware specifications):* 66.20 ms
- *Real-time factor (RTF) if model is causal:* N/A.
- *Training time (e.g. time per epoch):* 1h43m per epoch.
- *Memory footprint (e.g., memory usage during training, inference, model loading):* During Training:729MiB; during inference:293MiB;
- *Hardware specifications (CPU, GPU, TPU, memory capacity) used for training and inference:* Intel(R) Xeon(R) Bronze 3204 CPU @ 1.90GHz, 48GB of RAM, Tesla T4 GPU (16GB)
- *Number and type of GPUs used.* 1 NVIDIA Tesla T4 GPU (16GB)
- *Training process: Data augmentation, batch size, optimization algorithm, learning rate schedule (or other hyperparameter tuning details), number of training epochs, early stopping criteria (if any)*: No data augmentation was performed, batch size of 4 was used with the Adam optimizer and learning rate equals to 0.001, the learning rate was reduced by a factor of 0.5 on plateau with patience of 2, training was performed for 3 epochs.
- *Reproducibility:* Code available at: https://github.com/jrjoaorenato/recognavse-v2
- *Any Known limitations or constraints of the developed system:* The system can only process audio of the same context length as used on training, so it must be split and reunited for processing.
- *Any specific hardware or software requirements for running the system*: at least a 12GB GPU, the following packages should be installed: pytorch, numpy, tqdm, torchsummary, decord, librosa, scipy, pystoi, pesq, soundfile, torchvision, torchaudio, transformers, diffusers, timm.