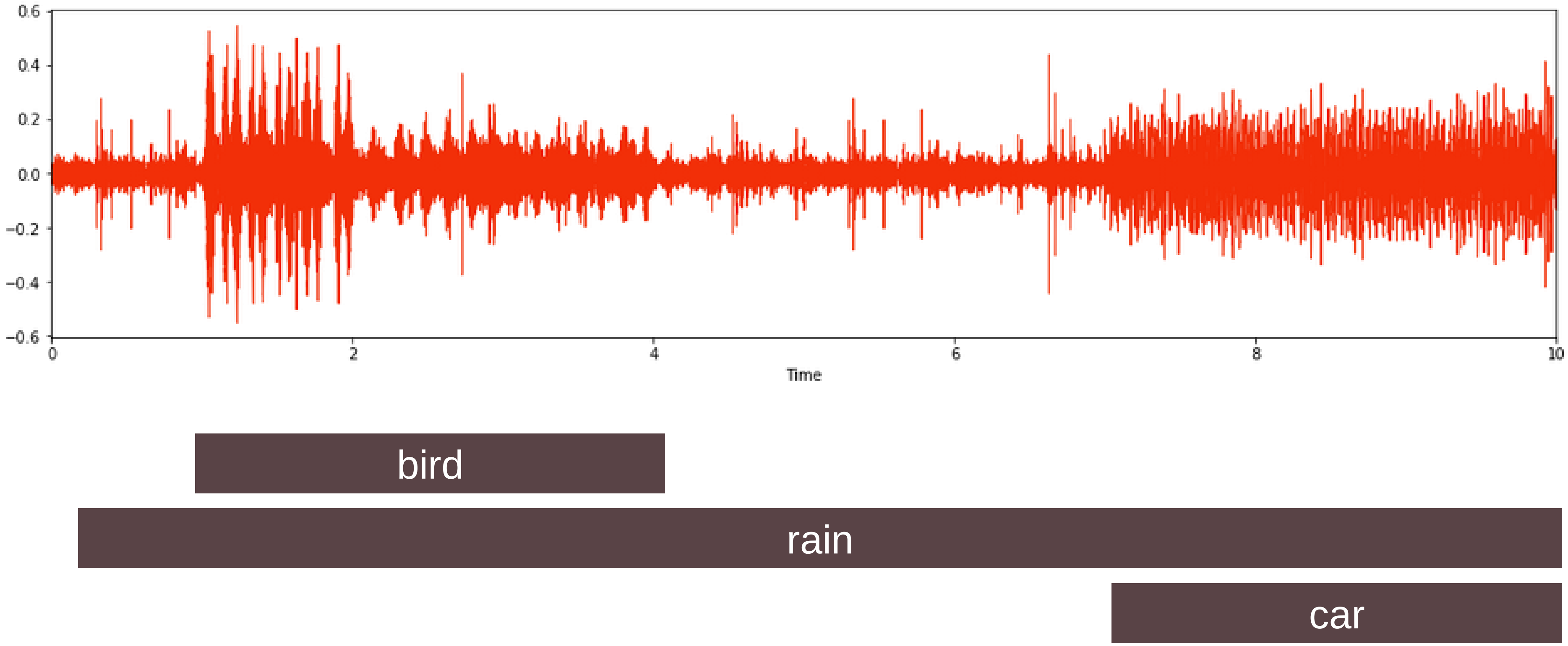
Sound events classification with CNN and data augmentation

Christophe Lesimple

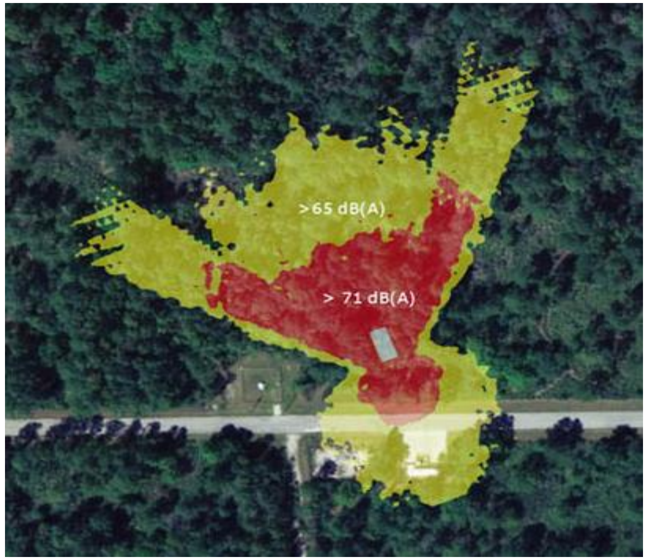
Sound Event Classification

- Source identification or event retrieval
- Sound event segmentation



Sound Classification: Applications

- Environmental sound/noise: qualitative measures [1]
- Medecine / Machine: diagnostic of pathologic sounds
- Hearing device: real time adjusment of amplification [2]



Hierarchy of classes

- Within Class [3]



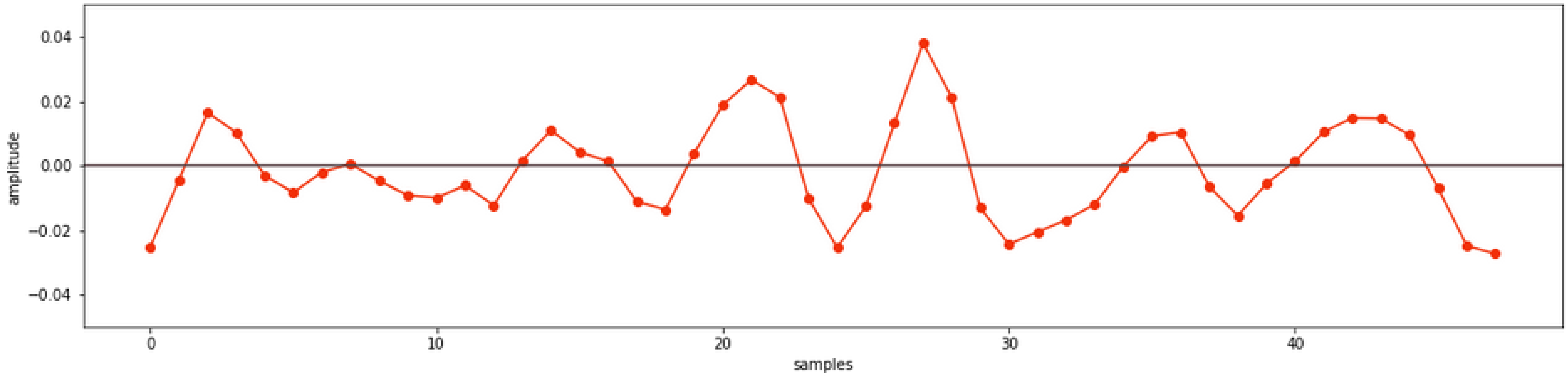

- Between Classes



- Combined Classes as soundscapes

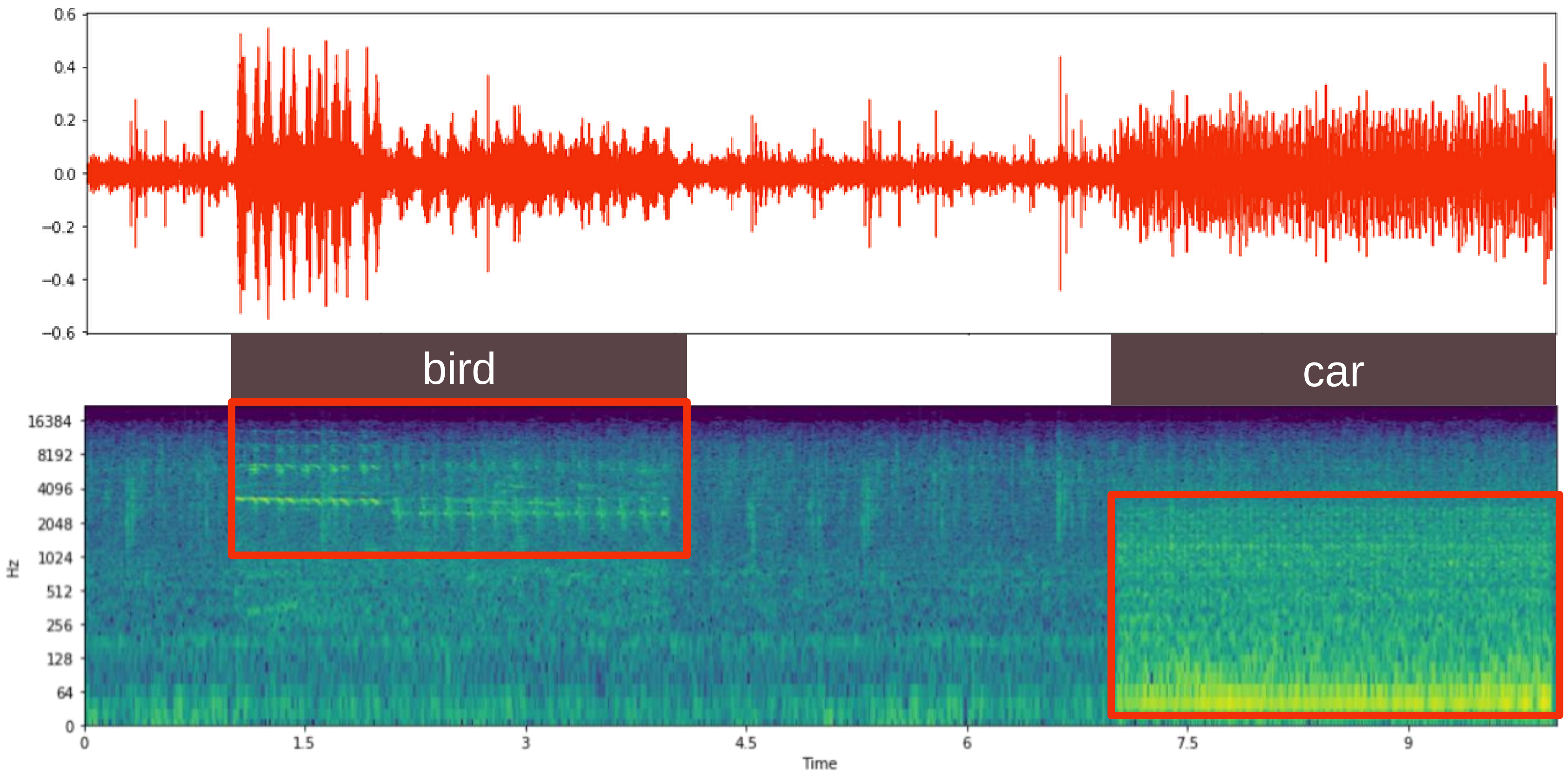

From a sound to 2D data

- Wavefile: amplitude variations over time
- Sampling frequency:
  - time resolution @ 44.1 kHz, 50 samples ~ 1.1 ms
  - influence the bandwith @ 44.1 kHz, fmax = 22.05kHz
  - large vectors without all the information



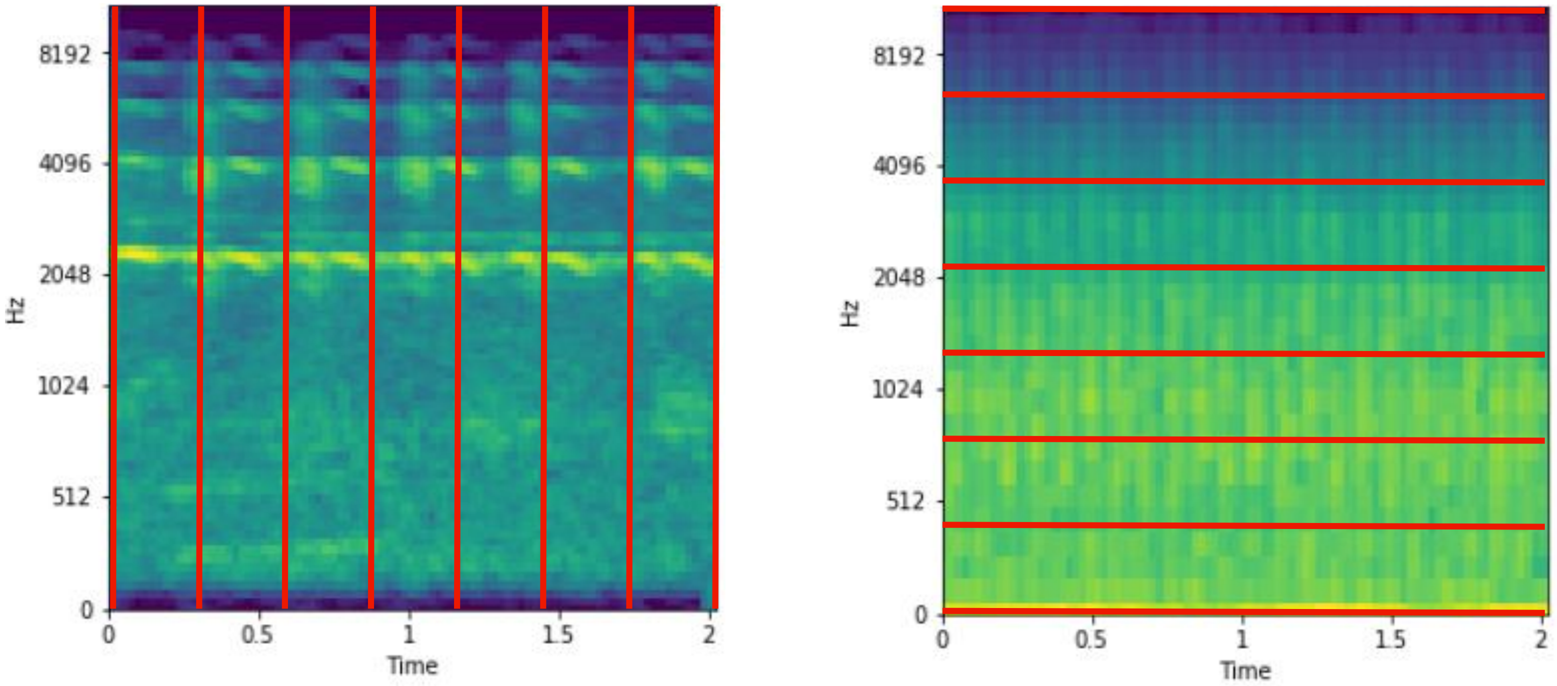
From sound to 3D data

- Convert the acoustic signal in time-frequency domain [4]:



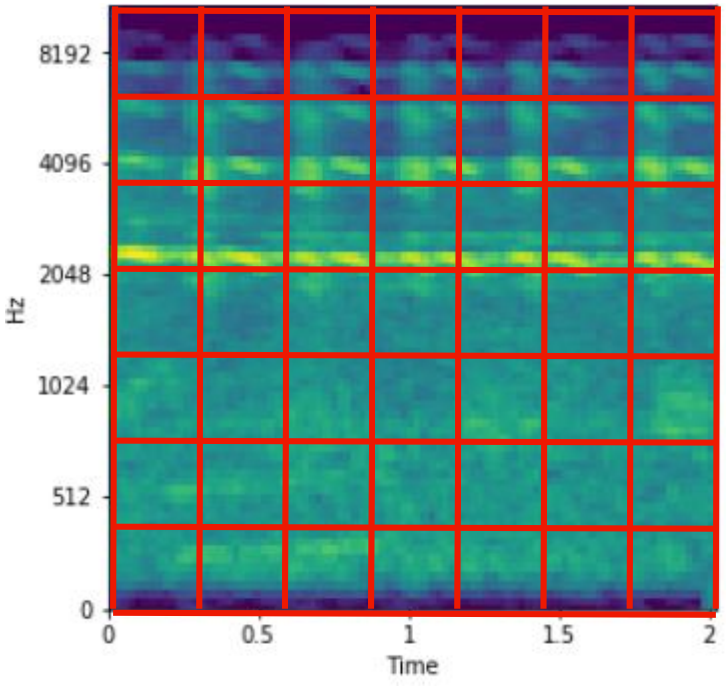
Data dimension vs. time/frequency

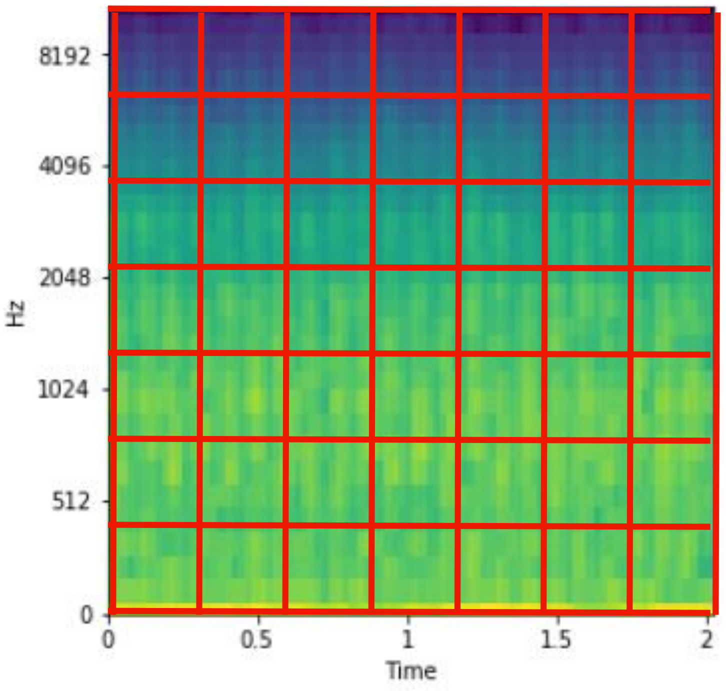
- Time resolution with `hop_length` → frames
- Frequency resolution with `n_mfcc` → bins



Data dimension vs. time/frequency

- Time resolution with `hop_length` → frames
- Frequency resolution with `n_mfcc` → bins





Adjust parameters based on sound source attributes e.g.:

- Modulation rate
- Frequency content

Same process for classification as image

Dataset from ESC 50

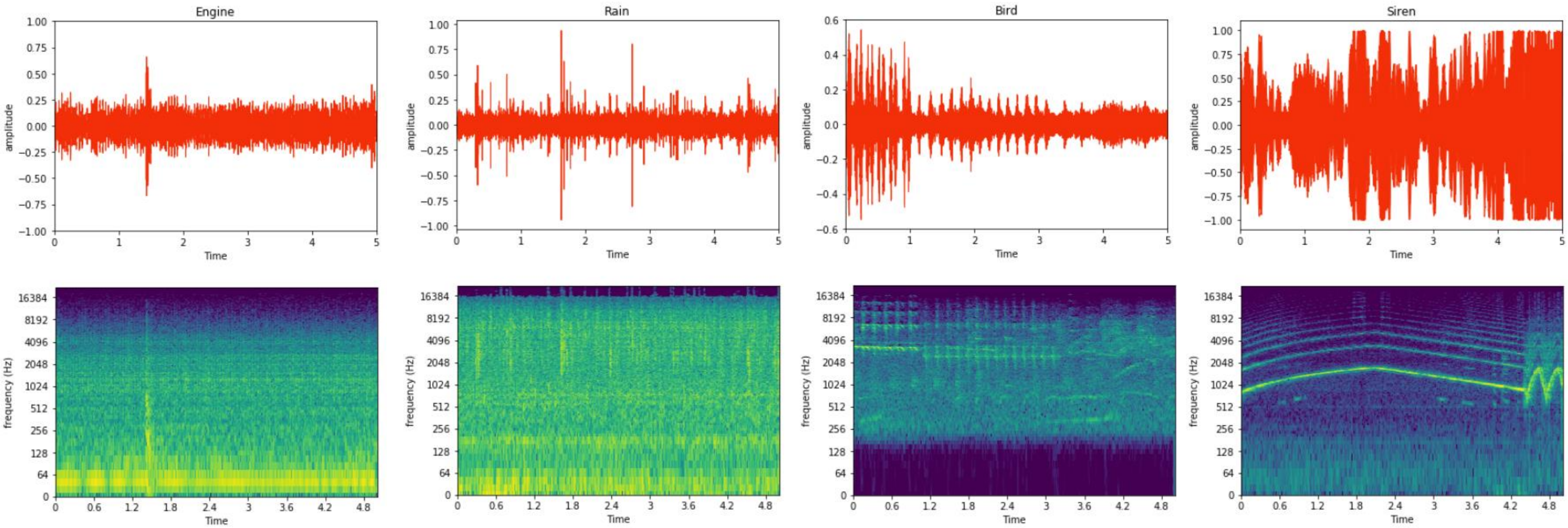
- Sounds from the freesound.org project, 5 seconds long
- Selection of 10 classes:
  - rain, sea waves, wind, crickets, birds,
  - car horn, train, siren, engine, church bells.

- Source [6] : github

- 40 samples per classes split in 80/20:
  - Training / validation set,
  - Test set.

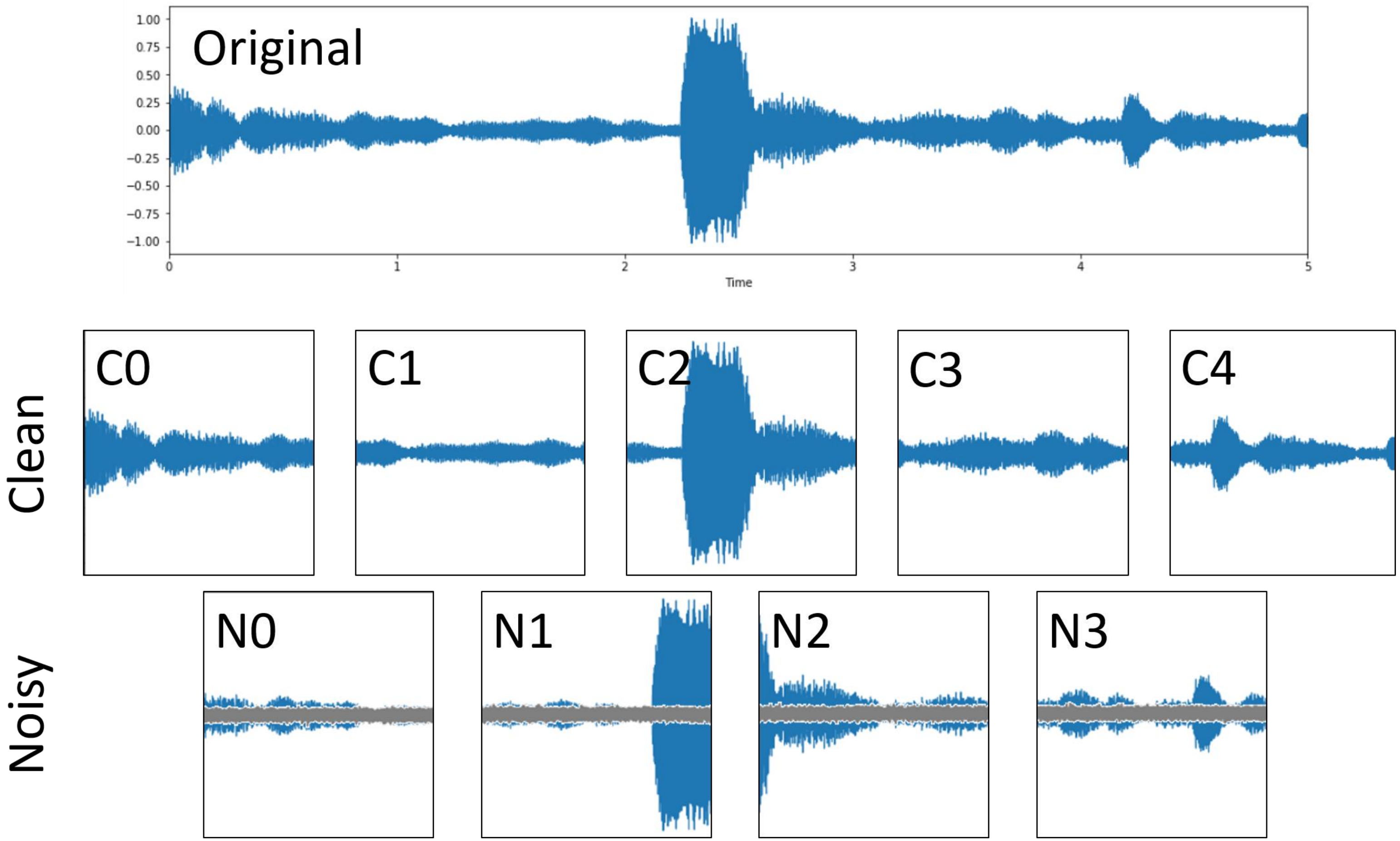
Supervised learning approach

- Features extracted from the wavefile
- Mapping between features and labels



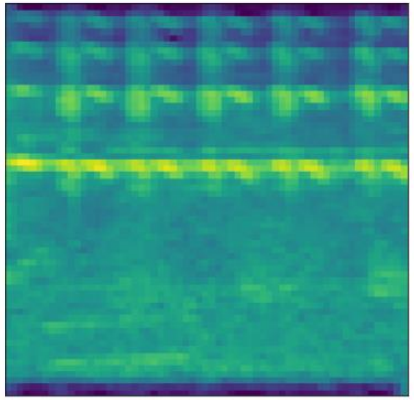
Data segmentation / augmentation

- 5s original file segmented in 9 files, 1s each,
- Data augmentation [7] by adding noise to each odd sample,
- Helps the CNN to see more relevant patterns at once and faster convergence.



Features dimensions

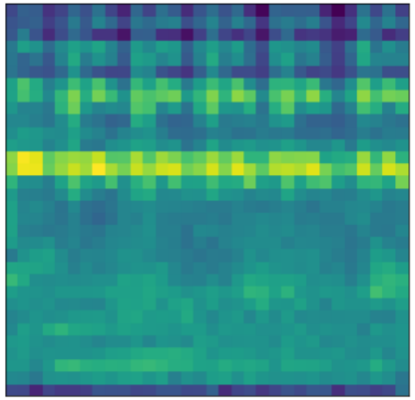
- Model 1: 63 x 63



- hop_length 512
- 63 MFCCs

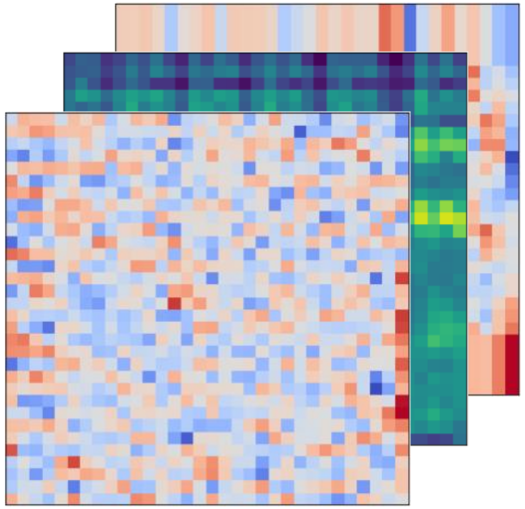
Mel-Frequency Cepstral Coefficient

- Model 2: 32 x 32



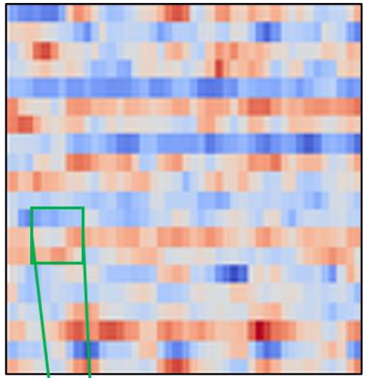
- hop_length 1024
- 32 MFCCs

- Model 3 [8]: 32 x 32 x 3

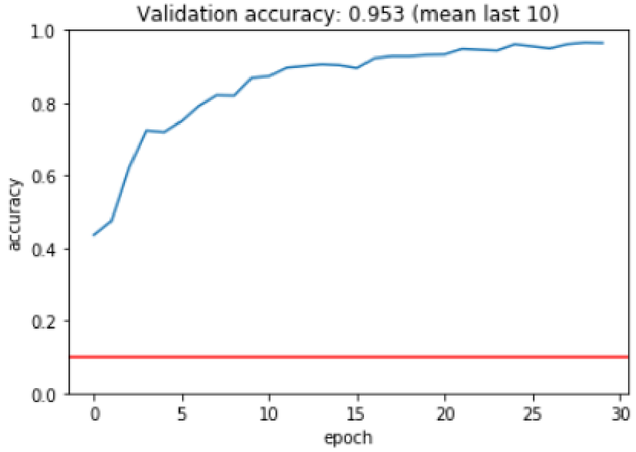


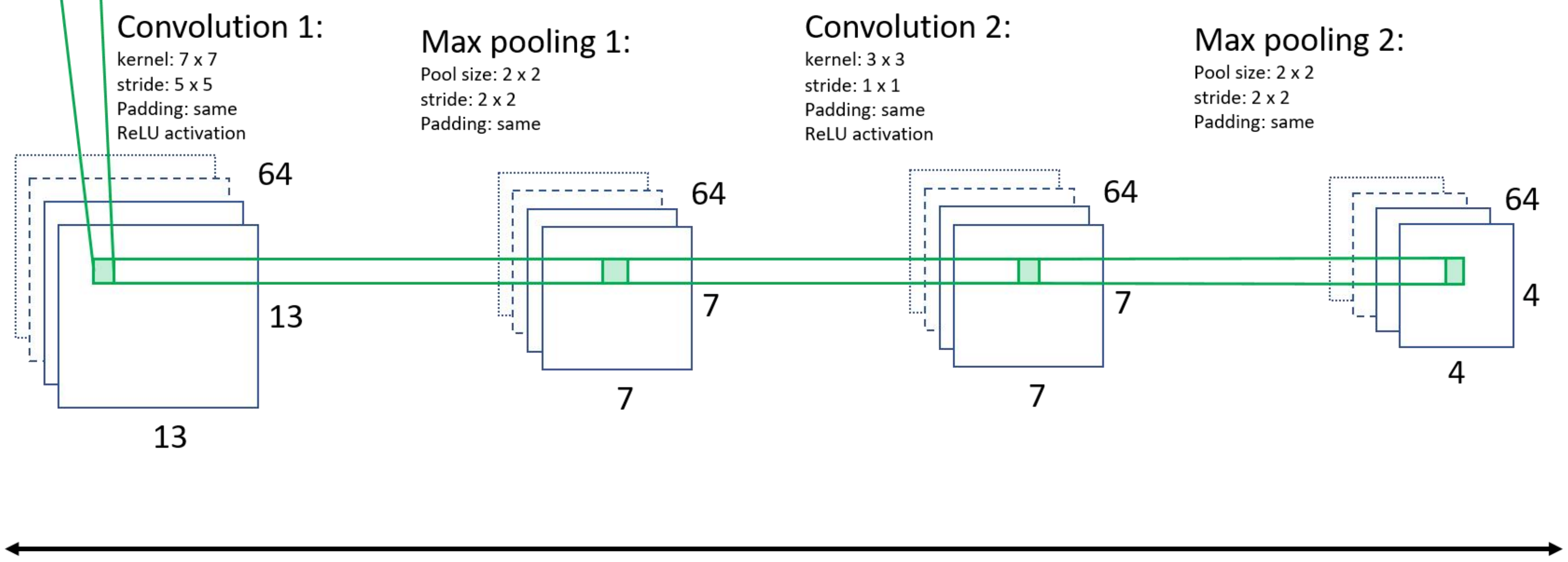
- Model 2 +
- Delta MFCCs
- Mel-spectrogram

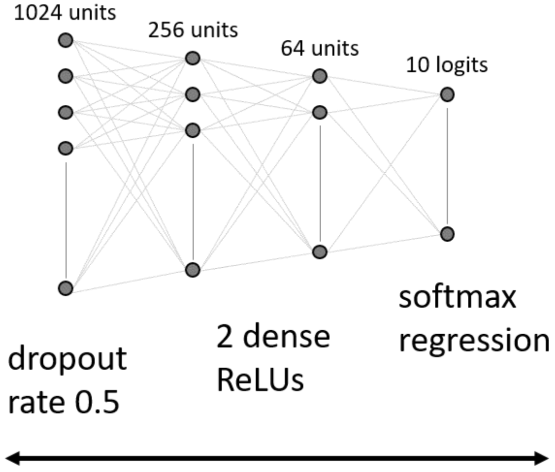
CNN with 2d convolution



Input: scaled MFCCs from selected 1s short sequence with data augmentation.
Dimensions 63 x 63 x 1



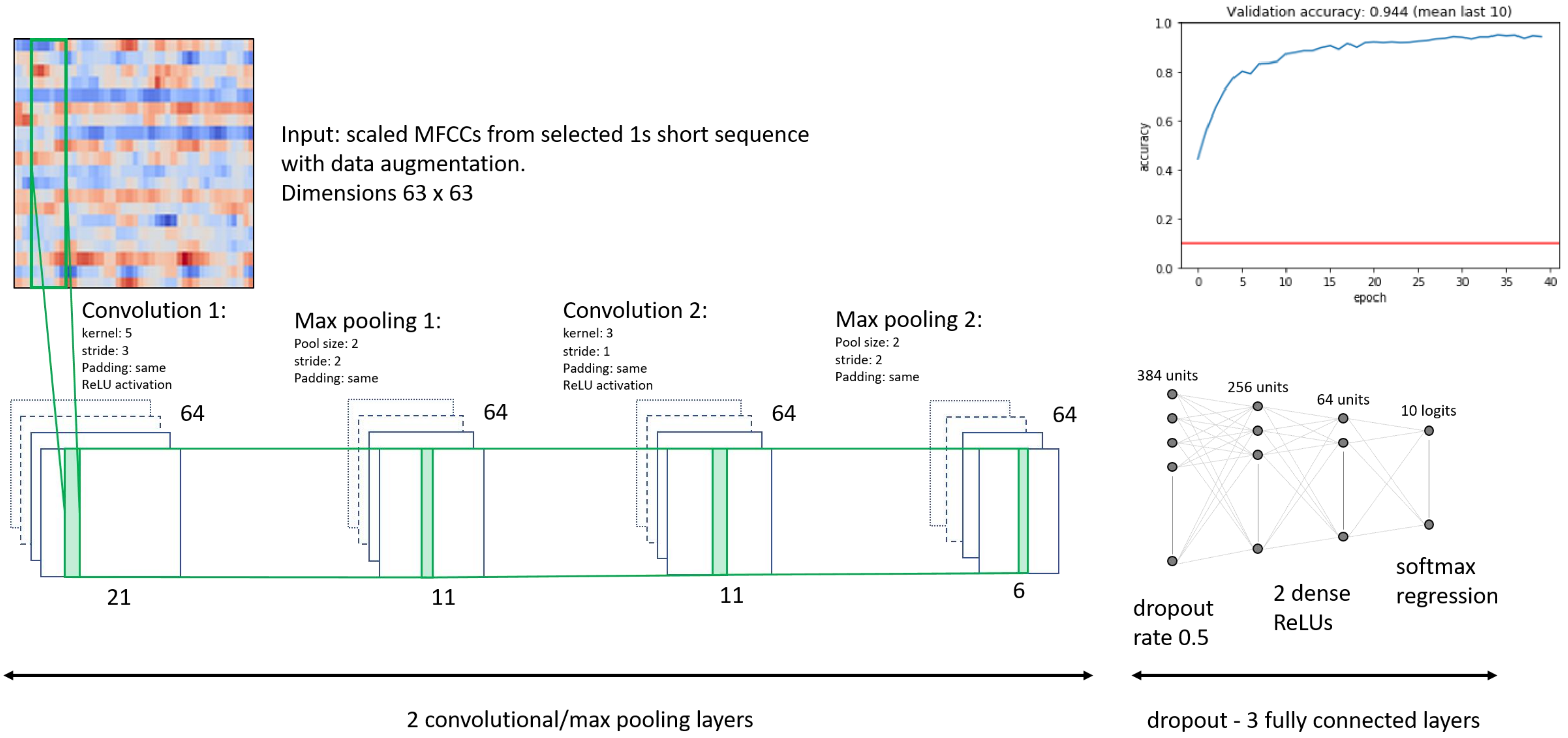




2 convolutional/max pooling layers

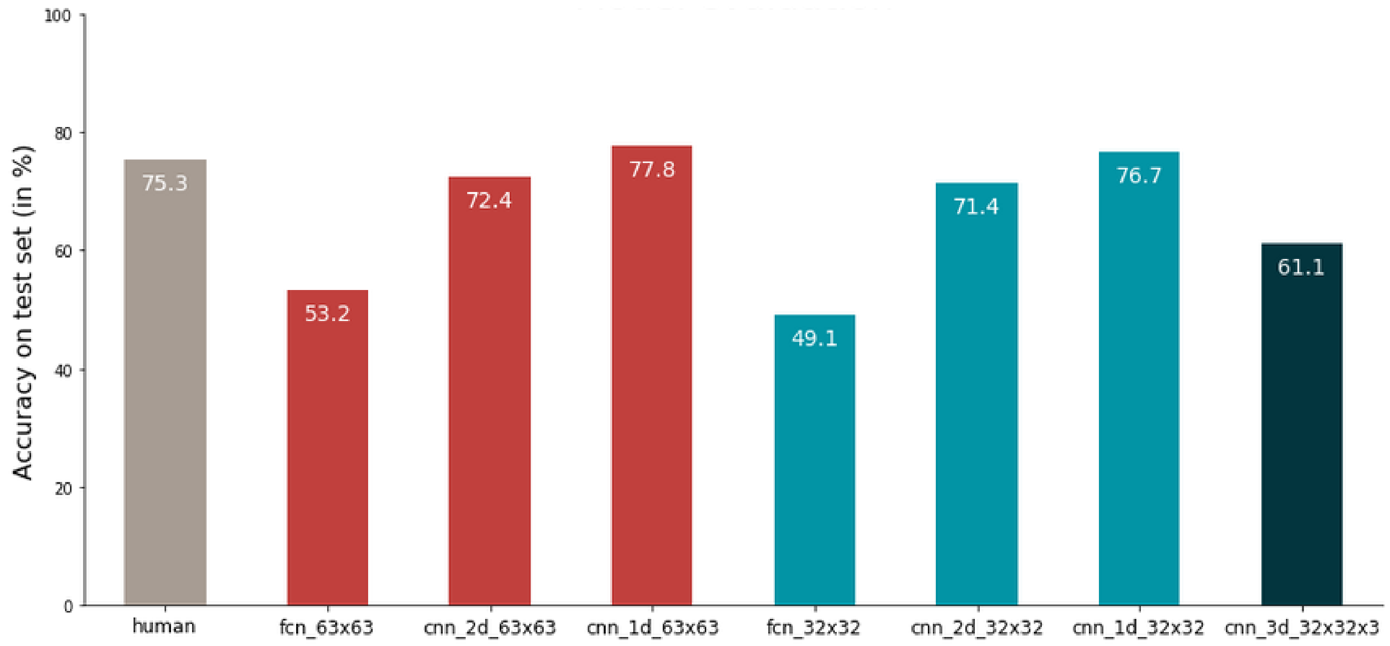
dropout - 3 fully connected layers

CNN with 1d convolution



Results with test set

- 1d convolution produces best results
- Reducing feature dimensions has a minor impact on accuracy
- Might be different for a within class classification task



References

1. Mijala et al. (2018), Environmental noise monitoring using source classification in sensors. Applied Acoustics, Volume 129, Pages 258-267.
2. Nordqvist, P. & Leijon, A. (2004), An efficient robust sound classification algorithm for hearing aids. J Acoust Soc Am. Jun;115(6):3033-41.
3. Xie et al. (2018), Acoustic classification of frog within-species and species-specific calls. Applied Acoustics, Volume 131, Pages 79-86.
4. Mitrovic, D., Zeppelzauer, M., & Breiteneder, C. (2010), Chapter 3 - Features for Content-Based Audio Retrieval. Advances in Computers, Elsevier, Volume 78, Pages 71-150.
5. Hershey et al. (2017), CNN Architectures for Large-Scale Audio Classification. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 131-135. 10.1109.
6. Piczak, K. (2015), ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015-1018, ACM.
7. Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, 24(3), 279-283.
8. Boddapati, V. et al. (2017), Classifying environmental sounds using image recognition networks. Procedia Computer Science, Volume 112, Pages 2048-2056.