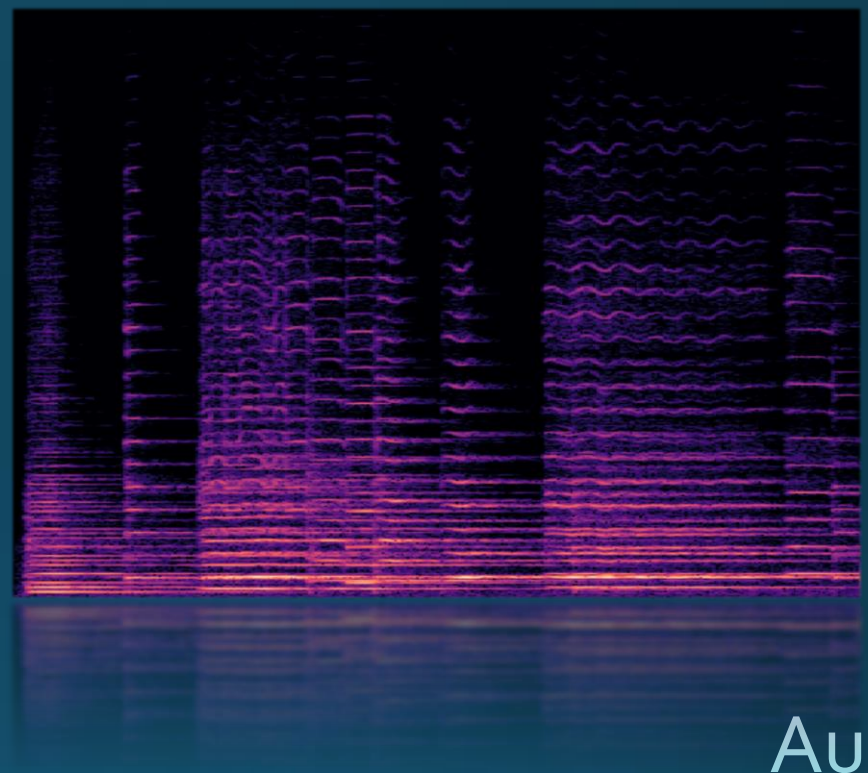


Austin

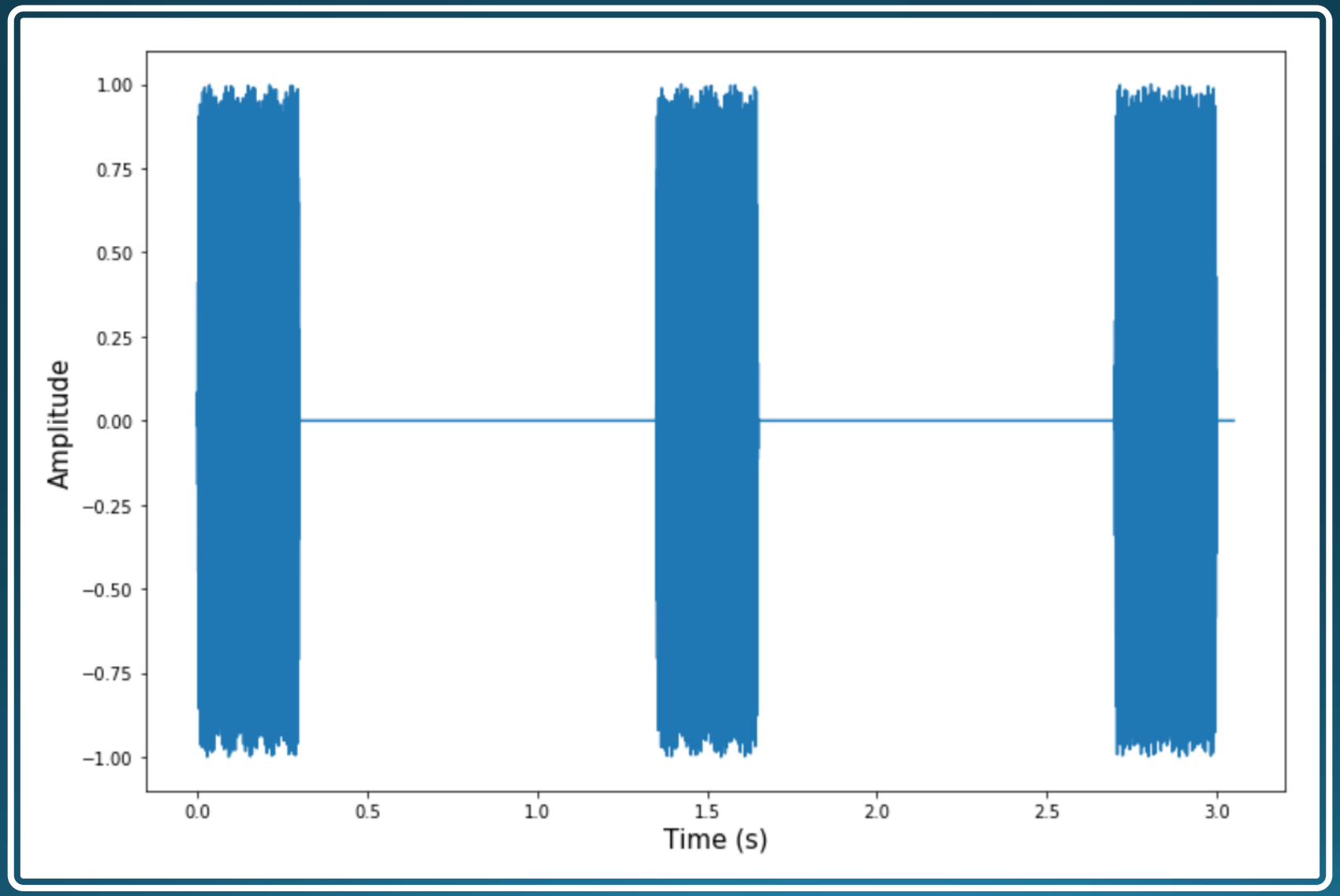
Sounds are hard.

UrbanSound8k Dataset

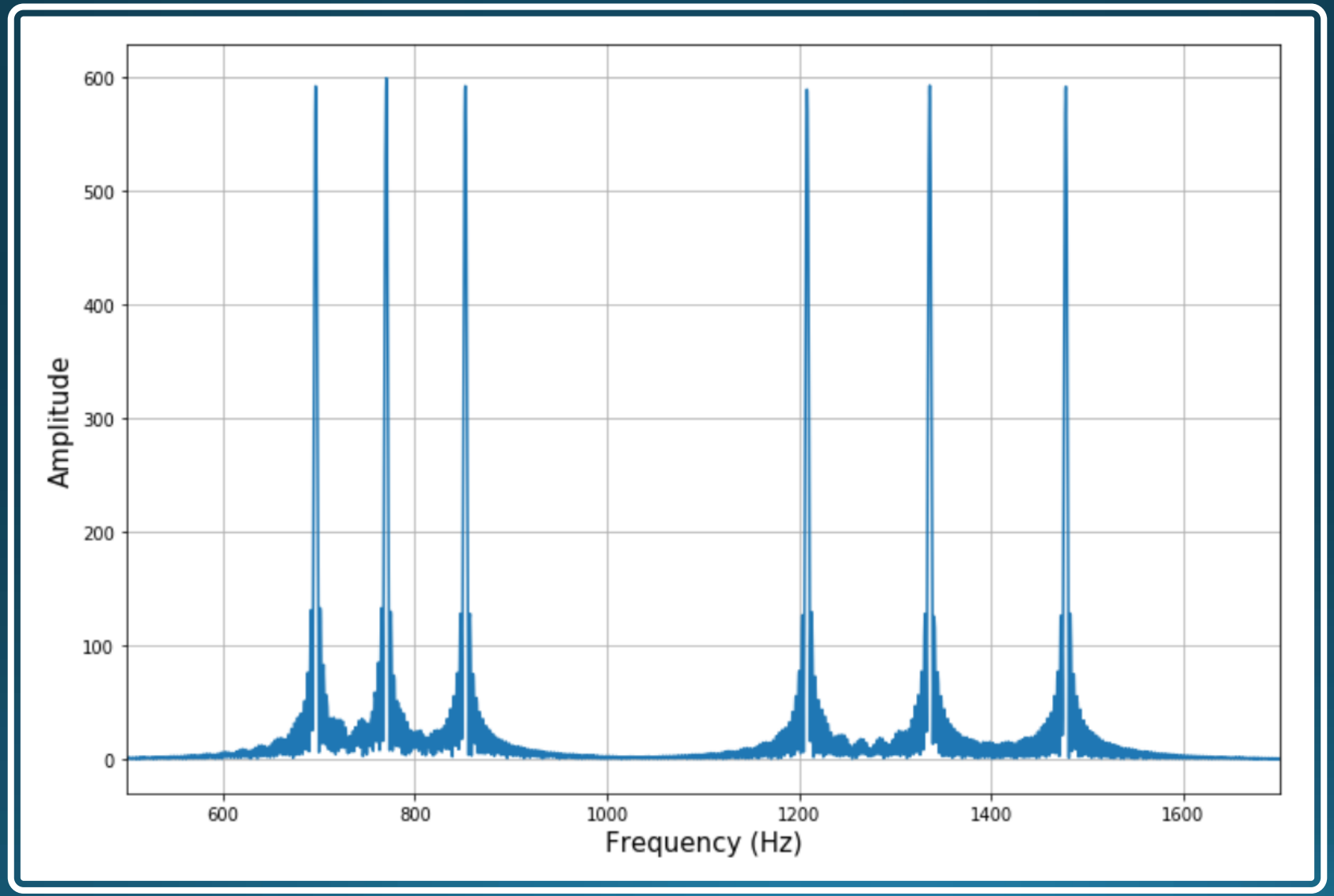
- 8732 wav files
- 10 classes
- Sound length from 1 to 4 seconds

Two Basic Representations of Sound

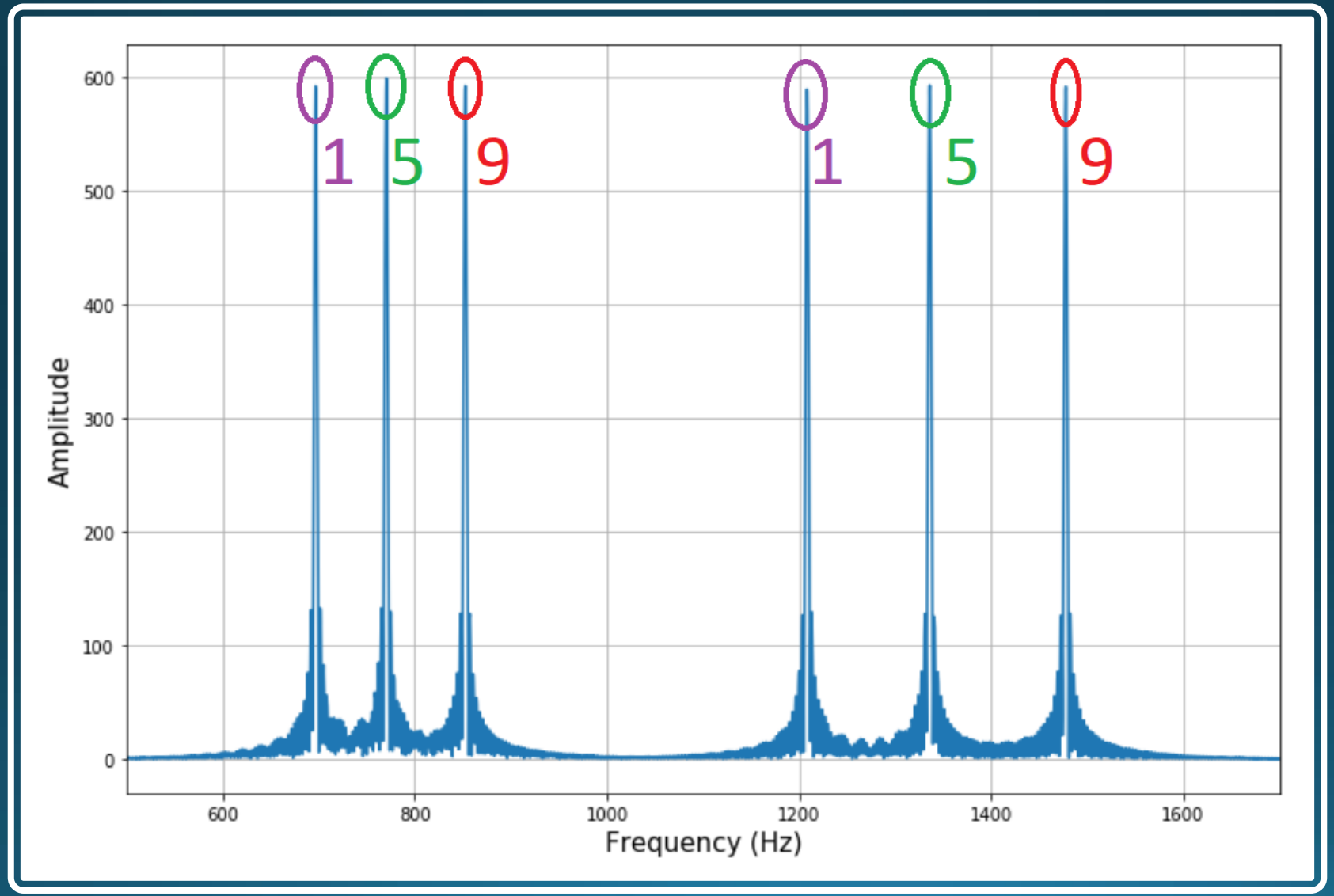
Time Representation



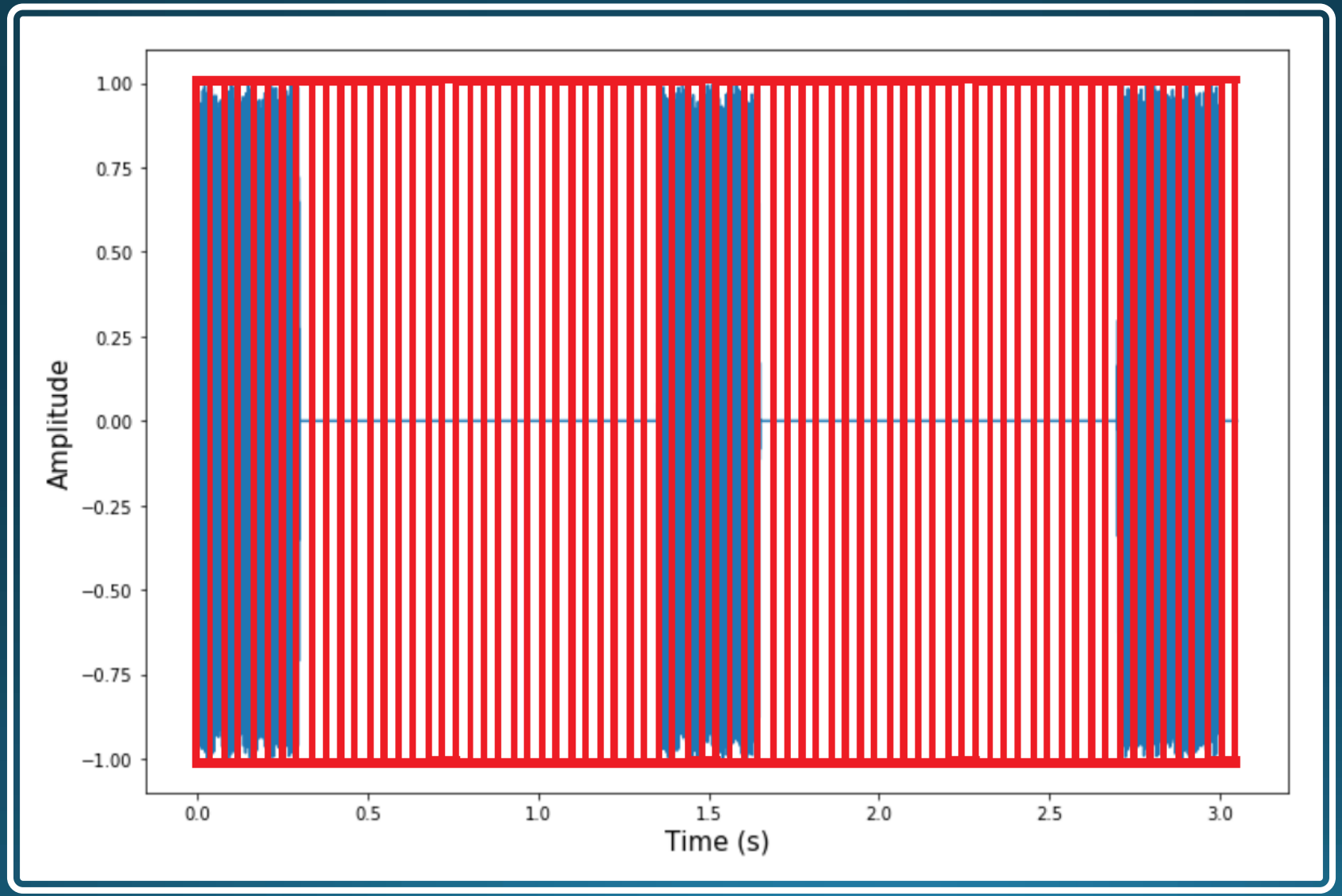
Frequency Representation



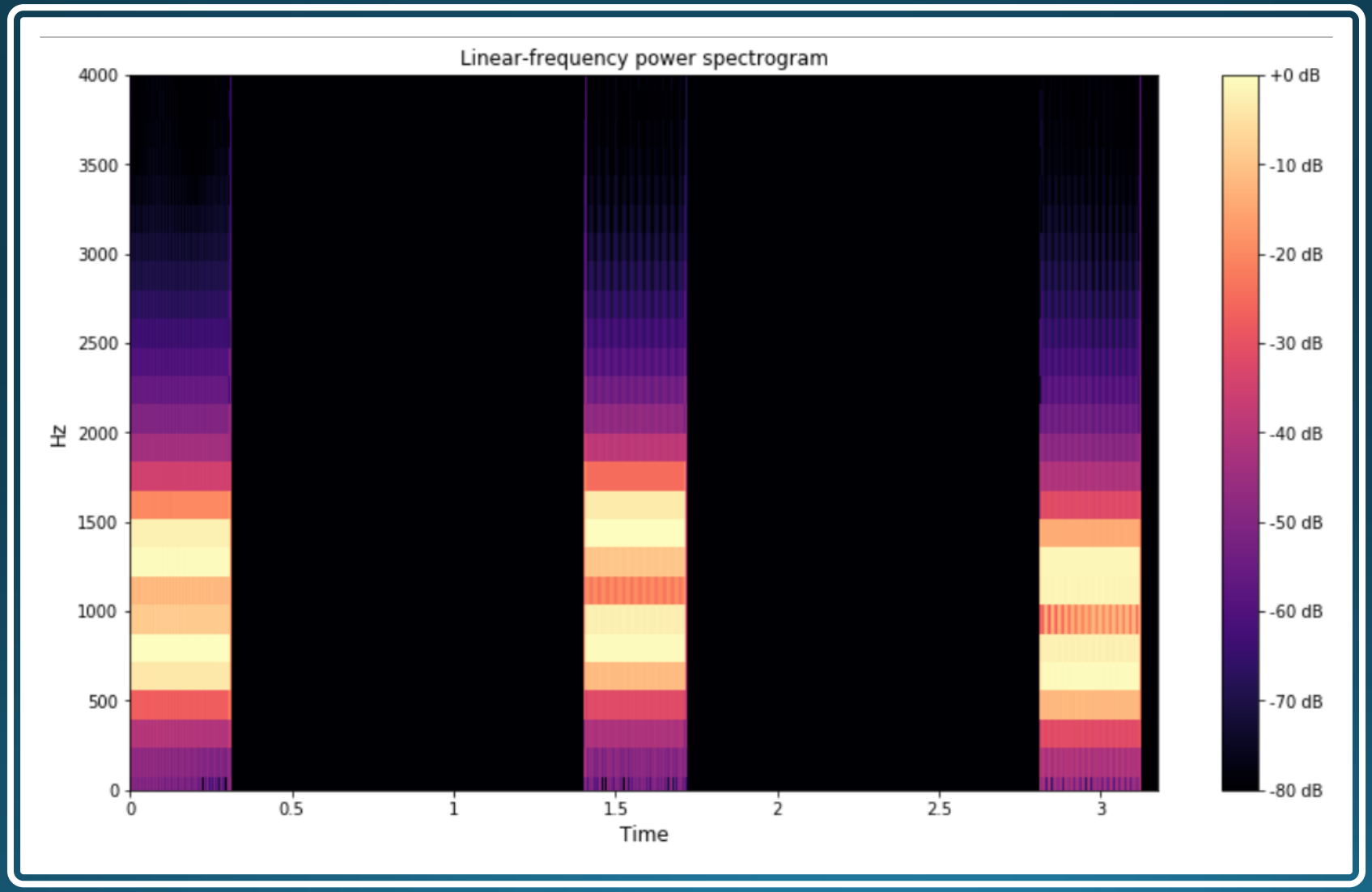
Frequency Representation



Clever Solution: STFT!

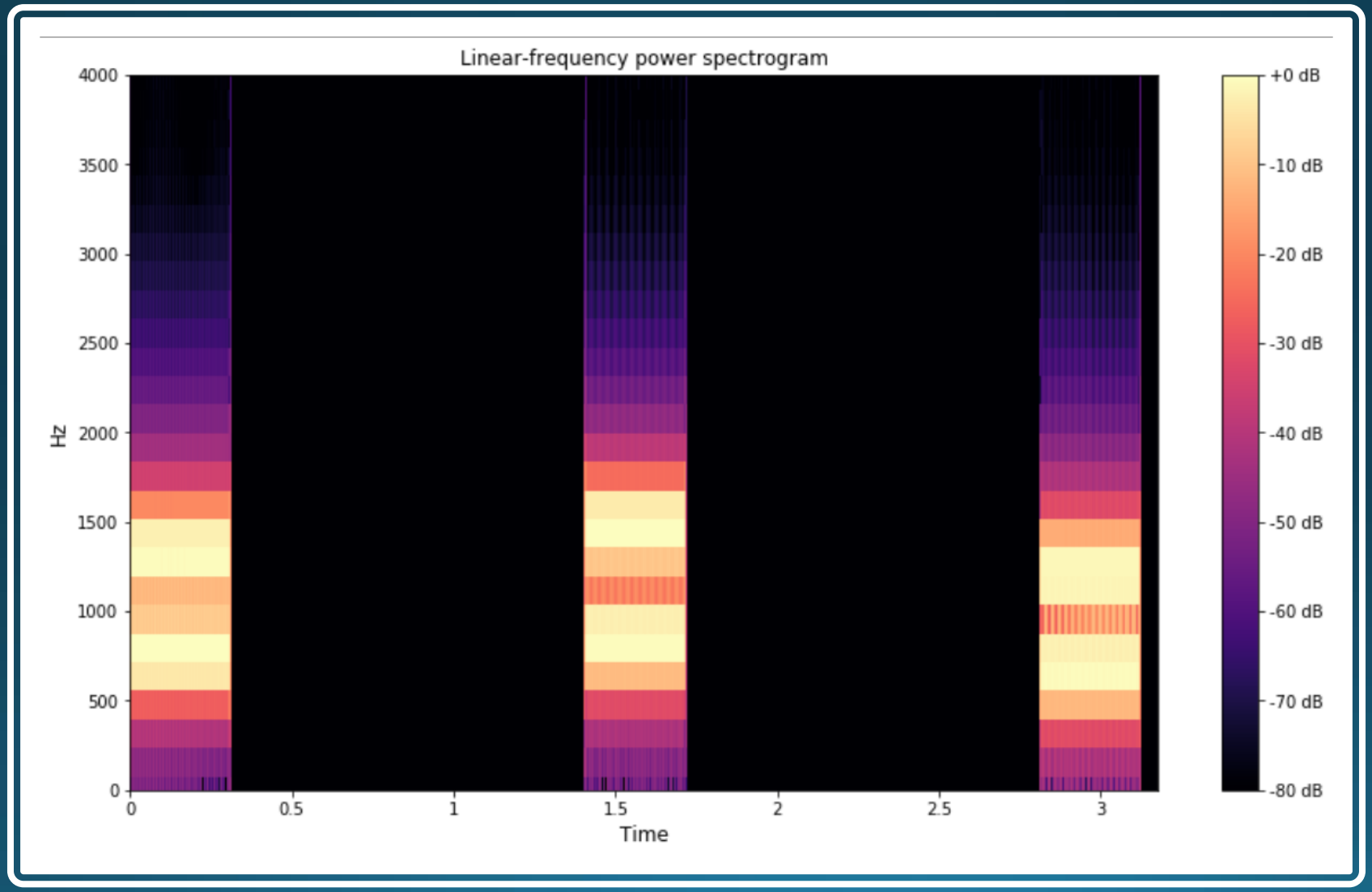


Spectrograms: Narrow Window

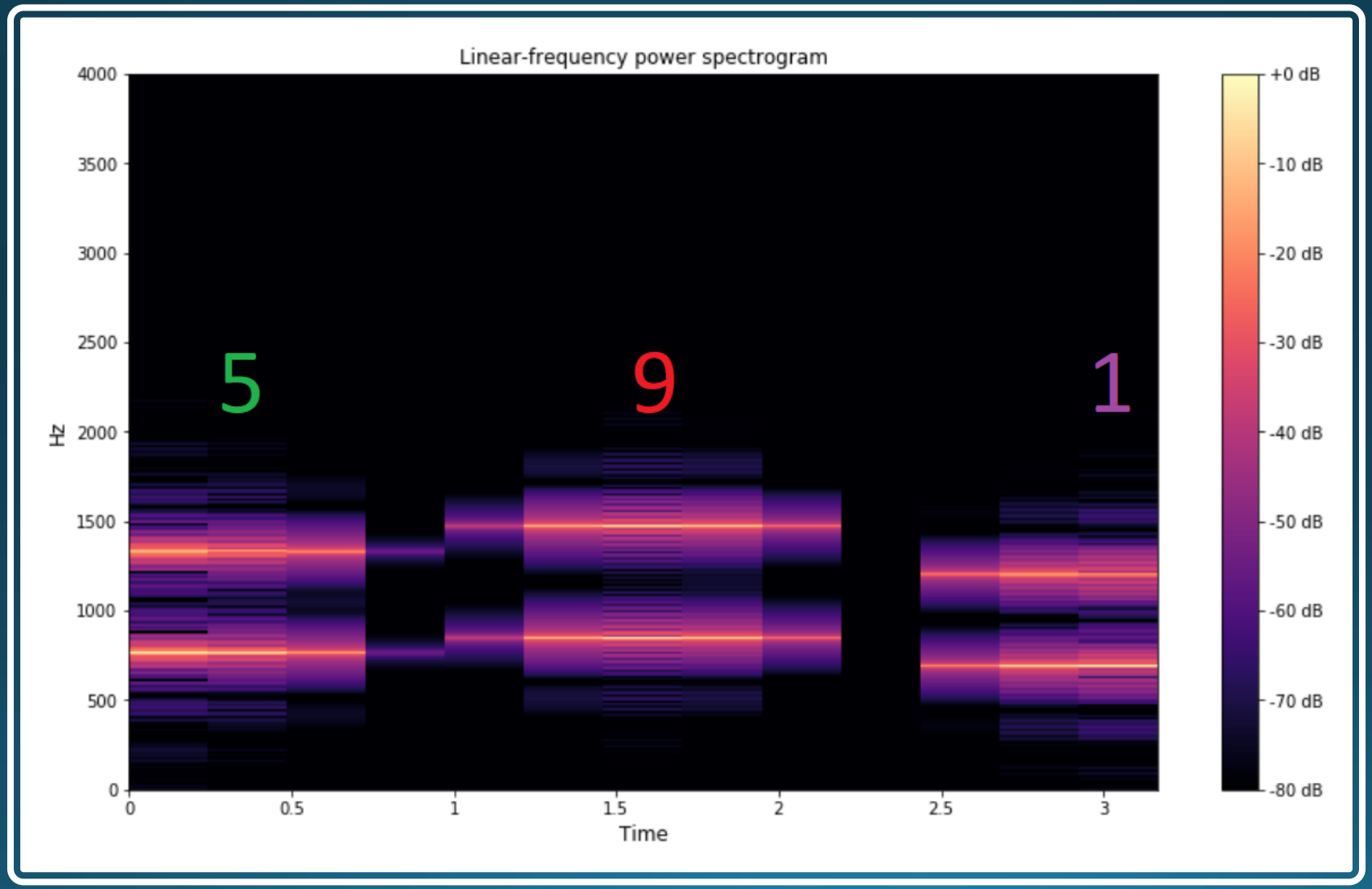


The Issue of Window Size

Spectrograms: Narrow Window

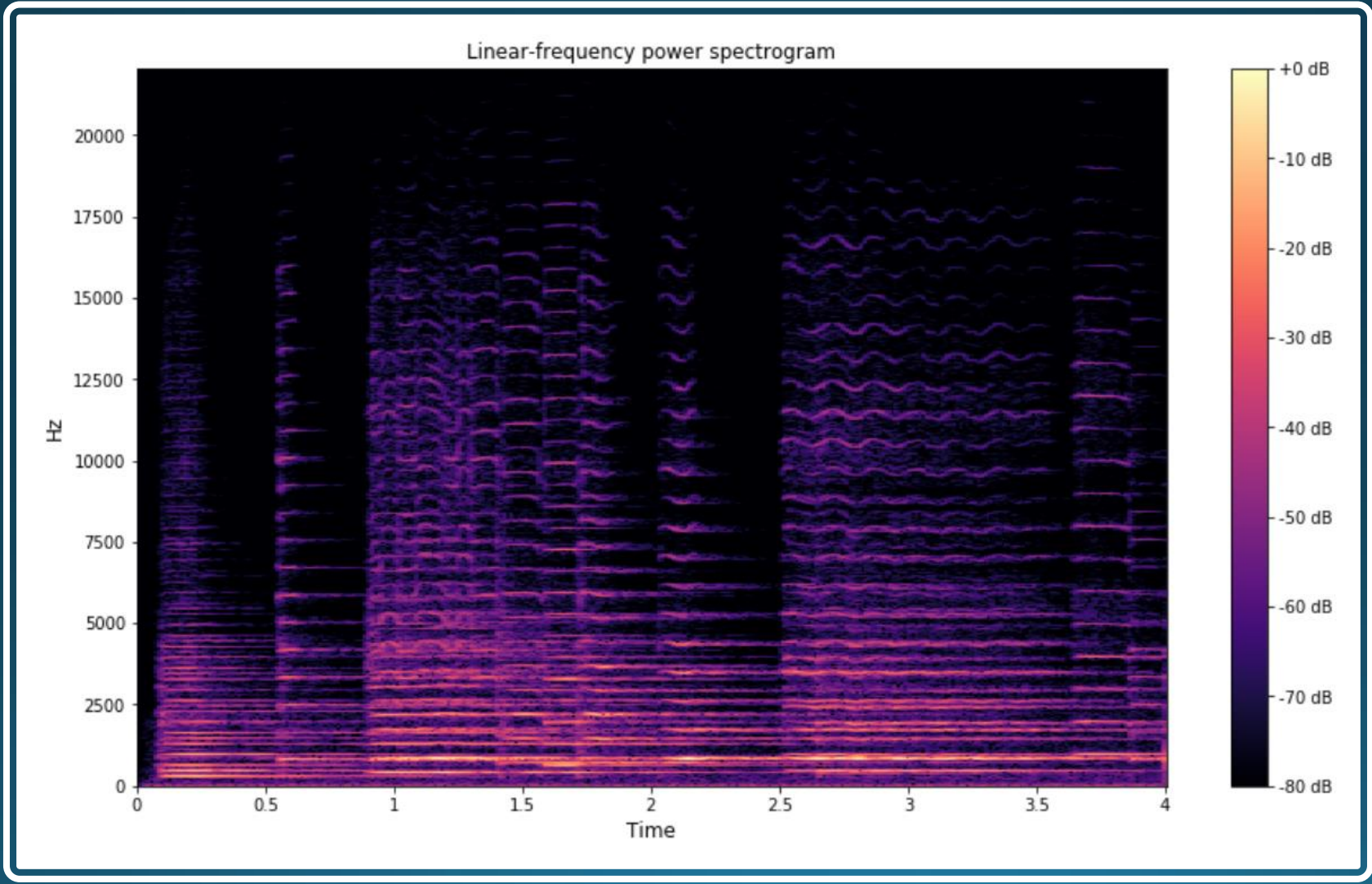


Wide Window

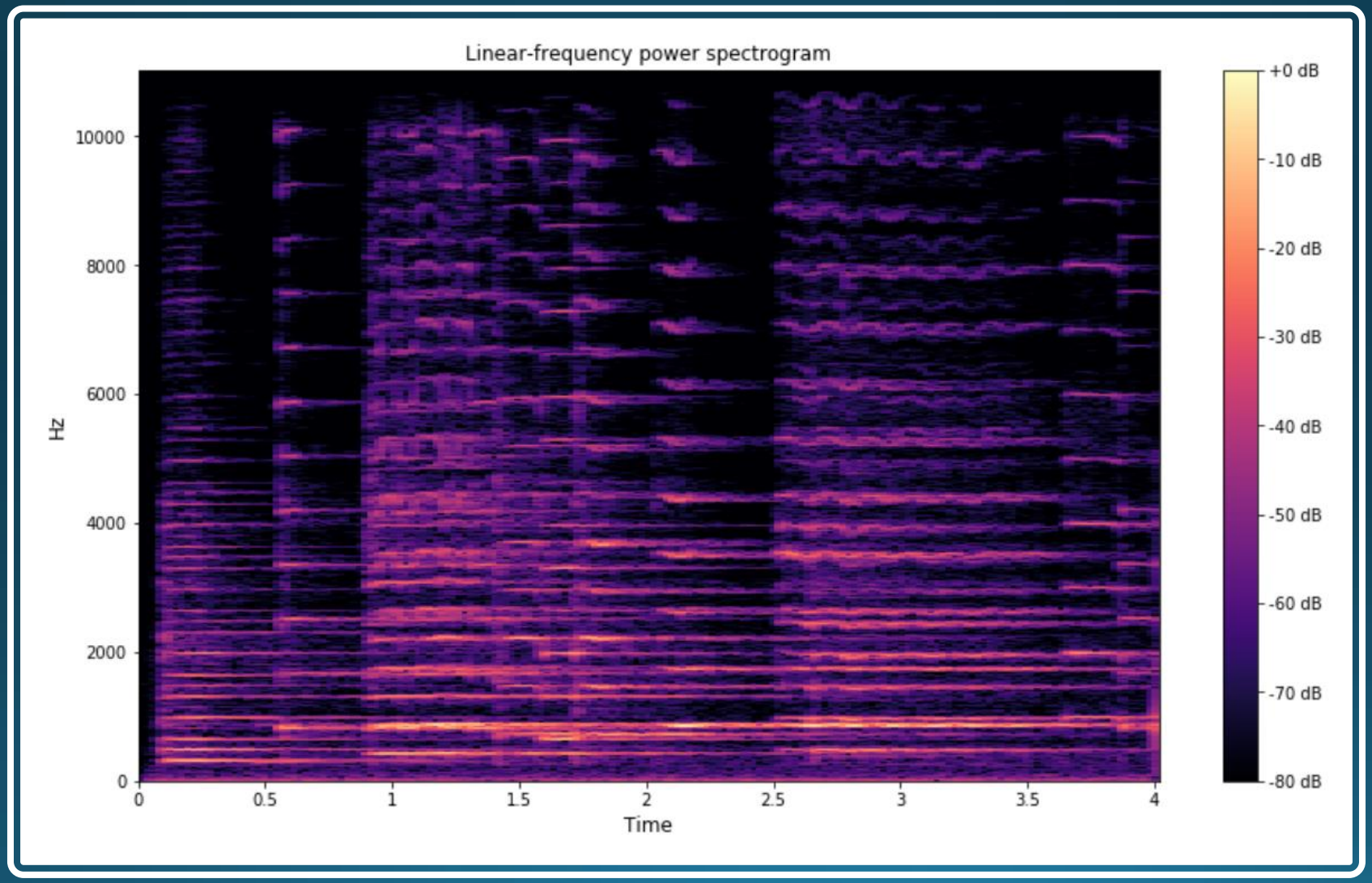

Linear-frequency power spectrogram
4000
3500
3000
2500
2000
1500
1000
500
0
Hz
5
9
1
0
0.5
1
1.5
2
2.5
3
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

The Issue of Samplerate

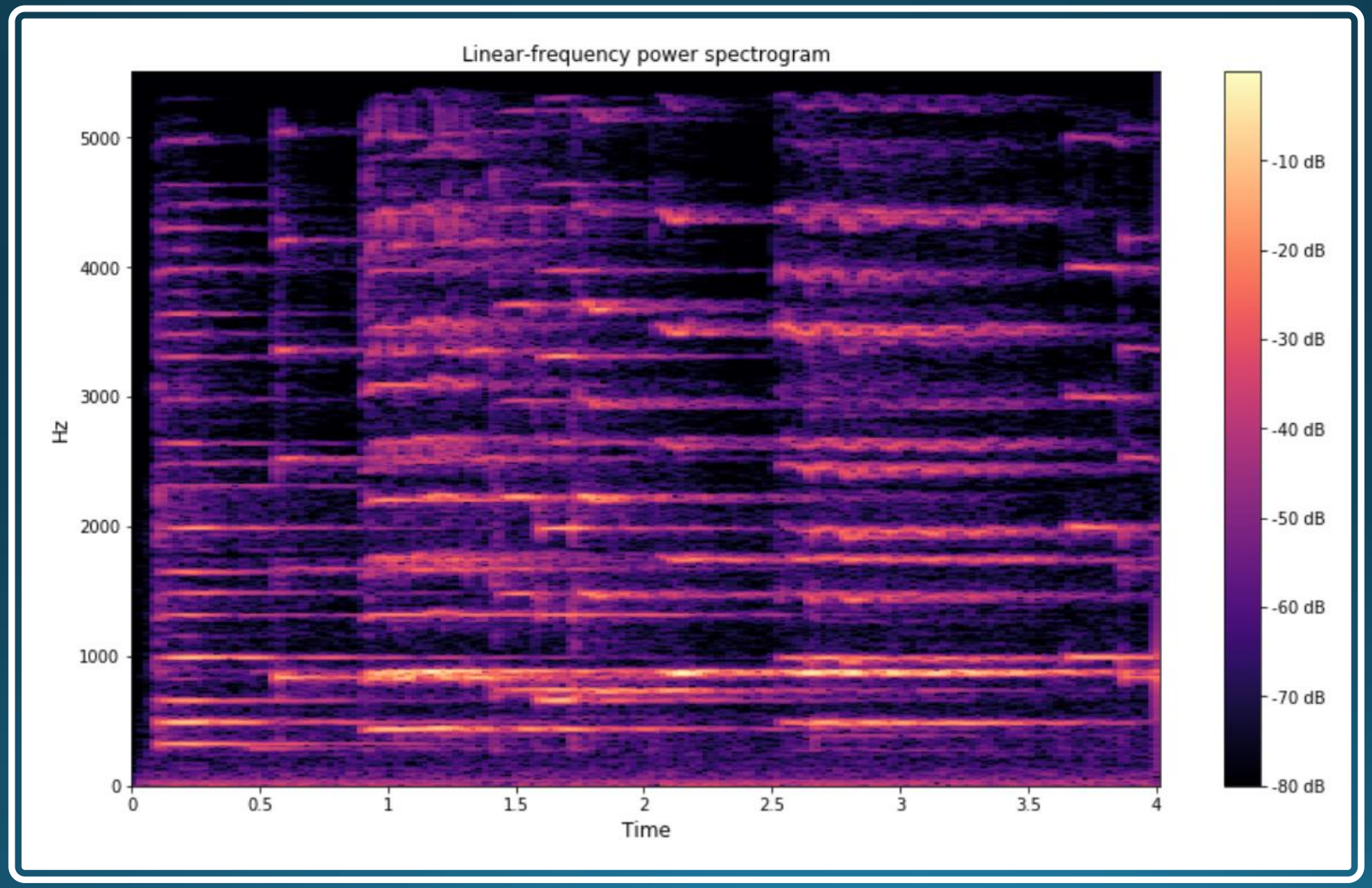
44100 Samplerate



22050 Samplerate

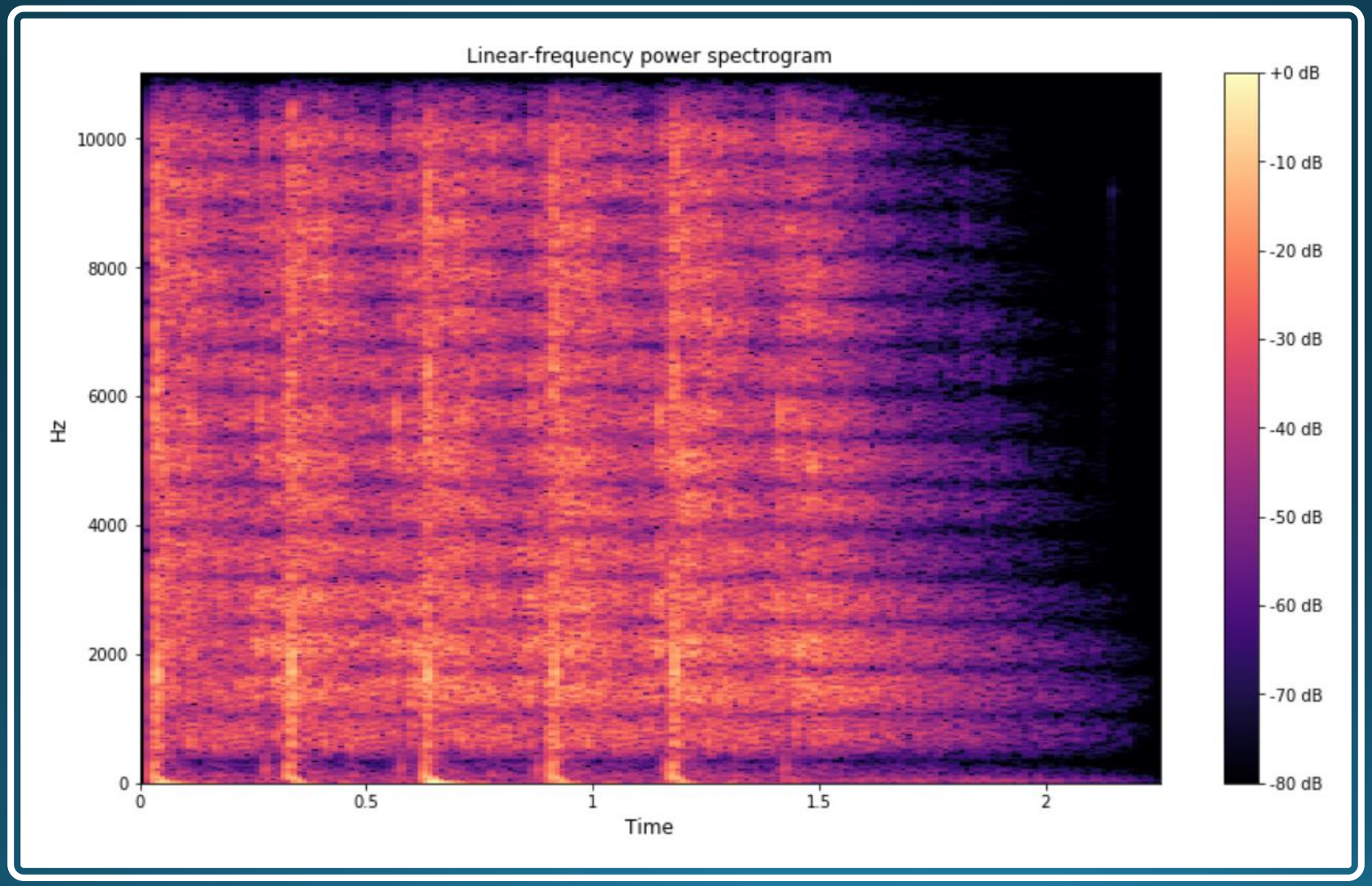


11025 Samplerate

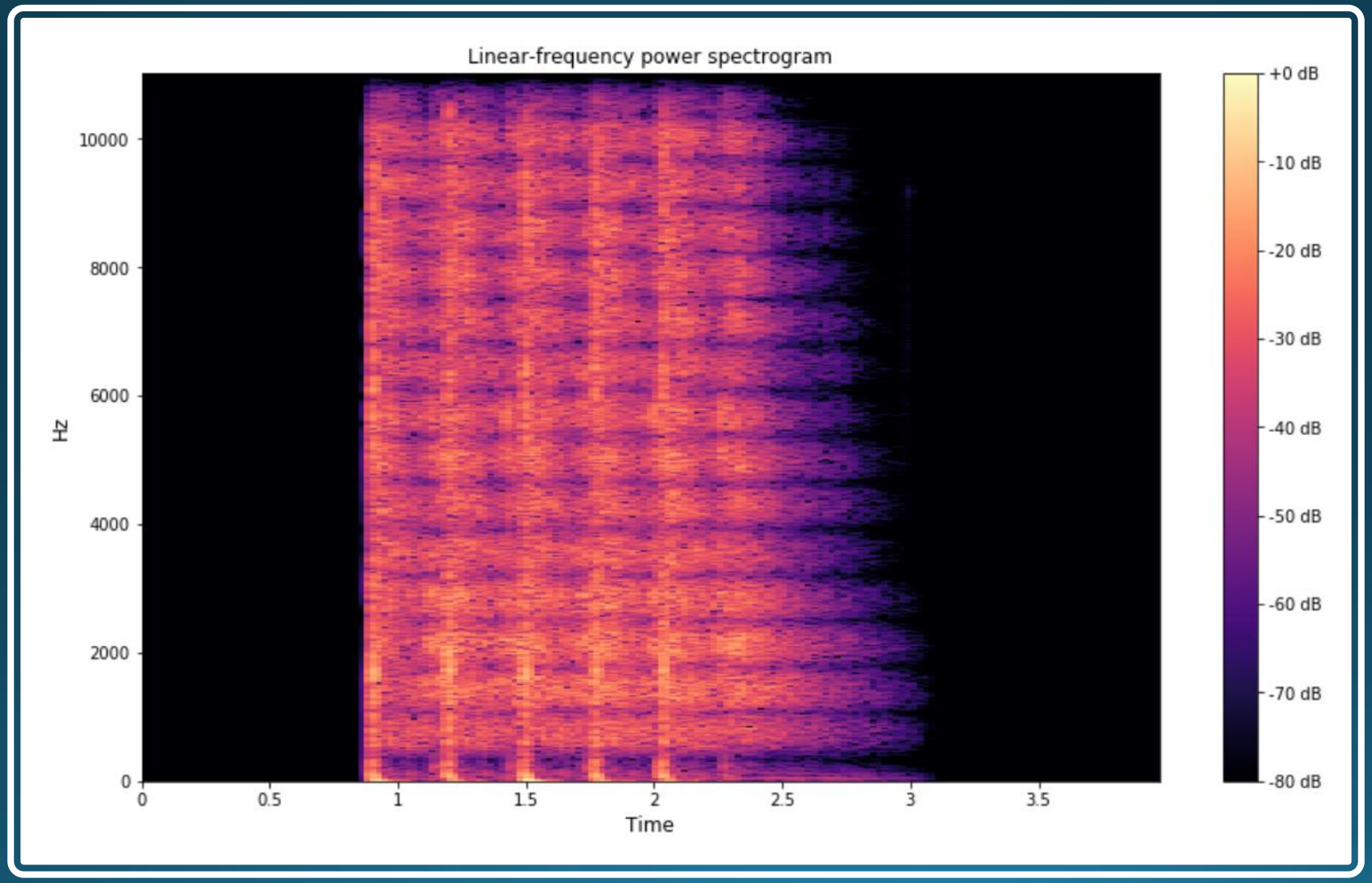


The Issue of Audio Length

Short Audio


Linear-frequency power spectrogram
10000
8000
6000
4000
2000
0
Hz
0
0.5
1
1.5
2
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Pad w/ Random Shift



The Issue of Questionable Data

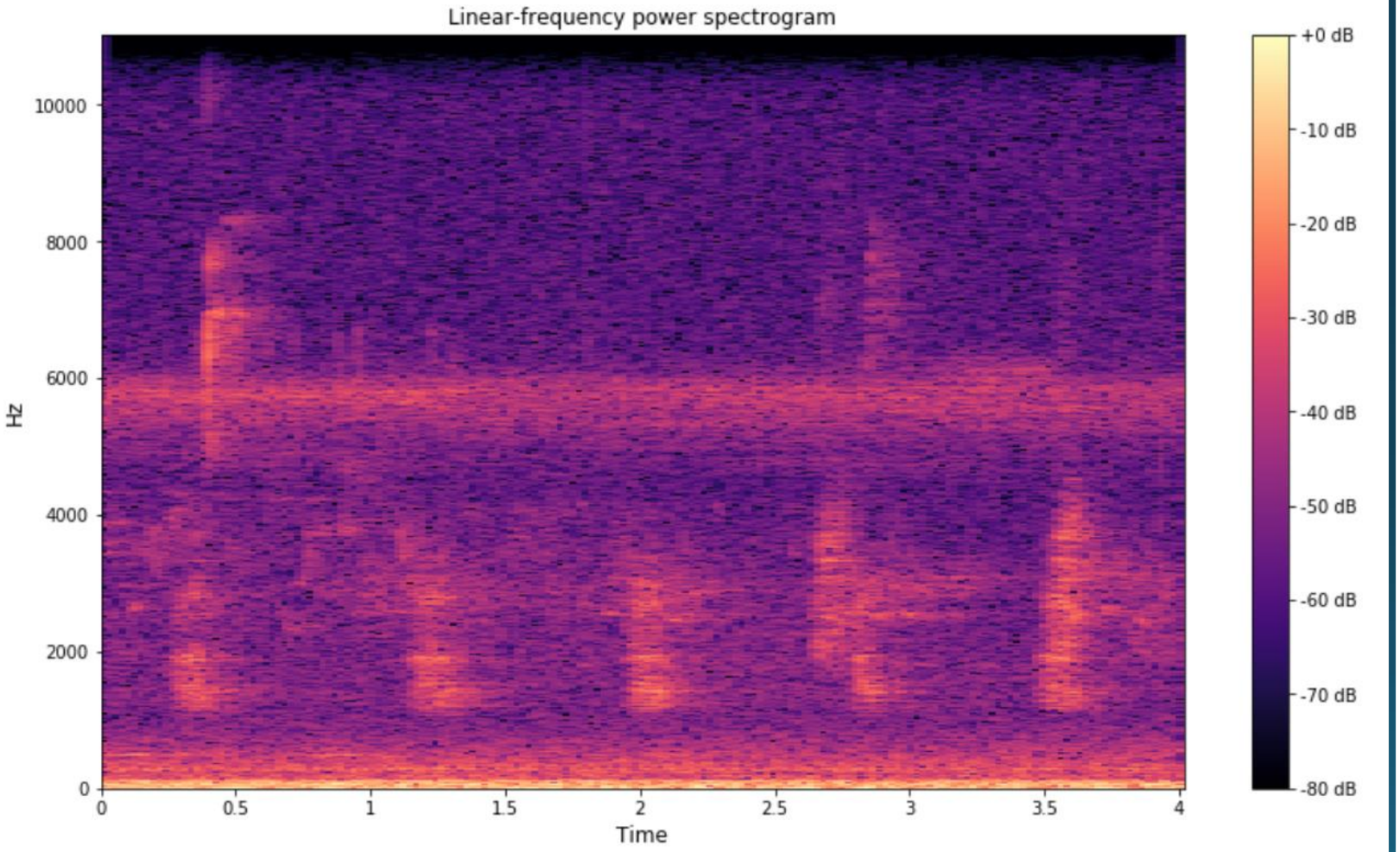
Guess that sound!

a) Music

b) Children

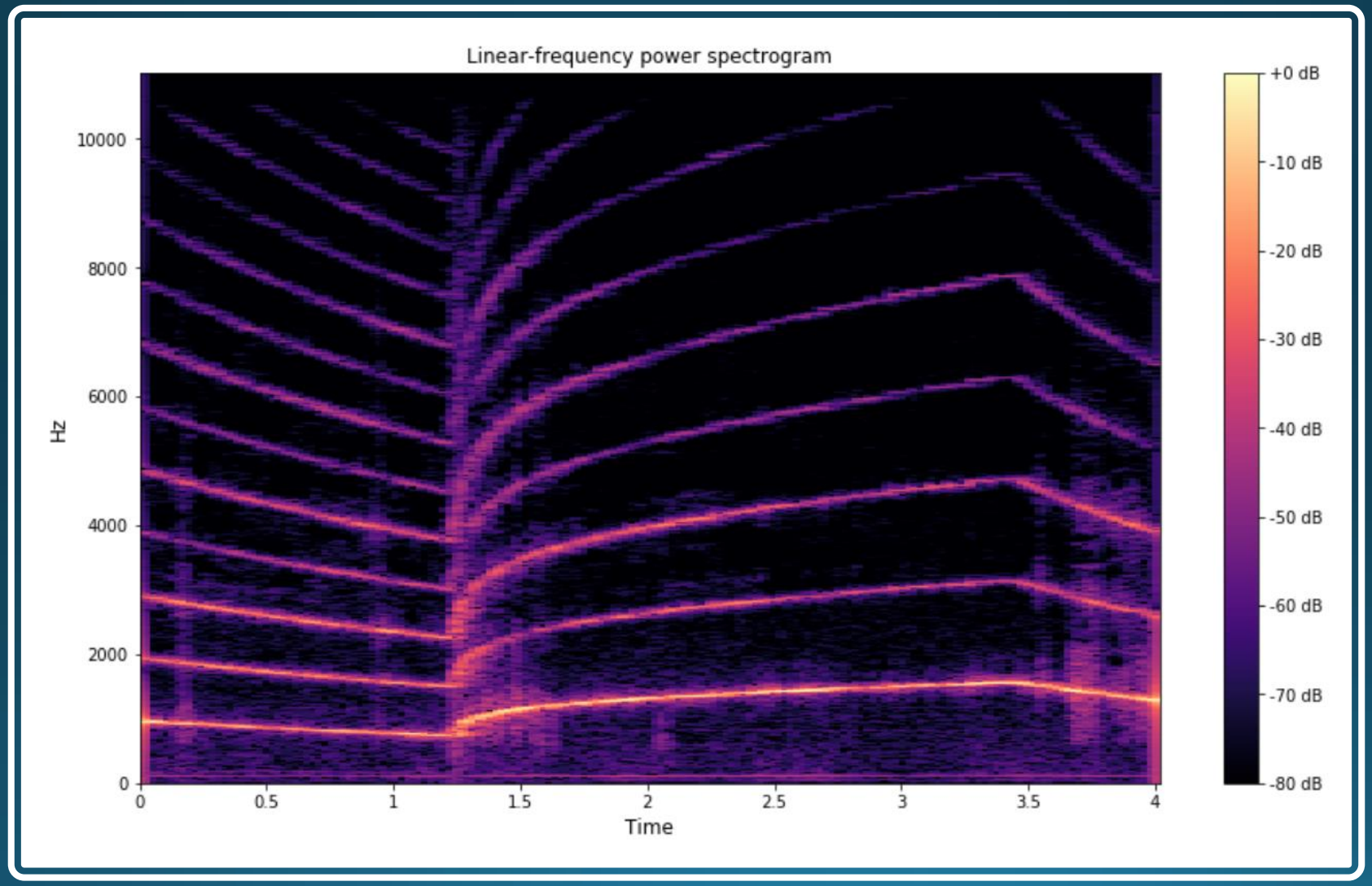
c) AC

d) Drill

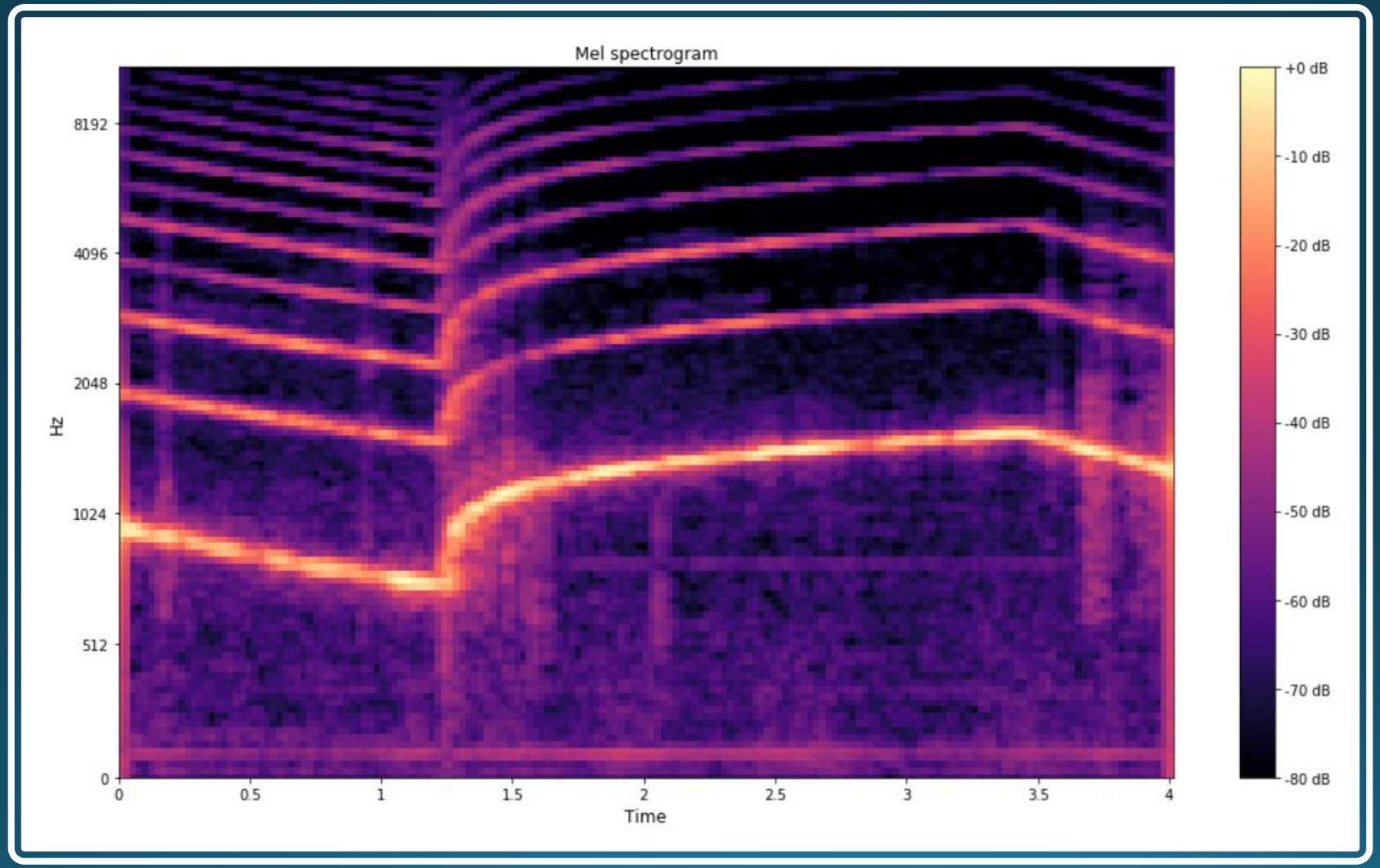


The Issue of Input Dimensions

Siren: 1025 x 101



Siren: 128 x 101

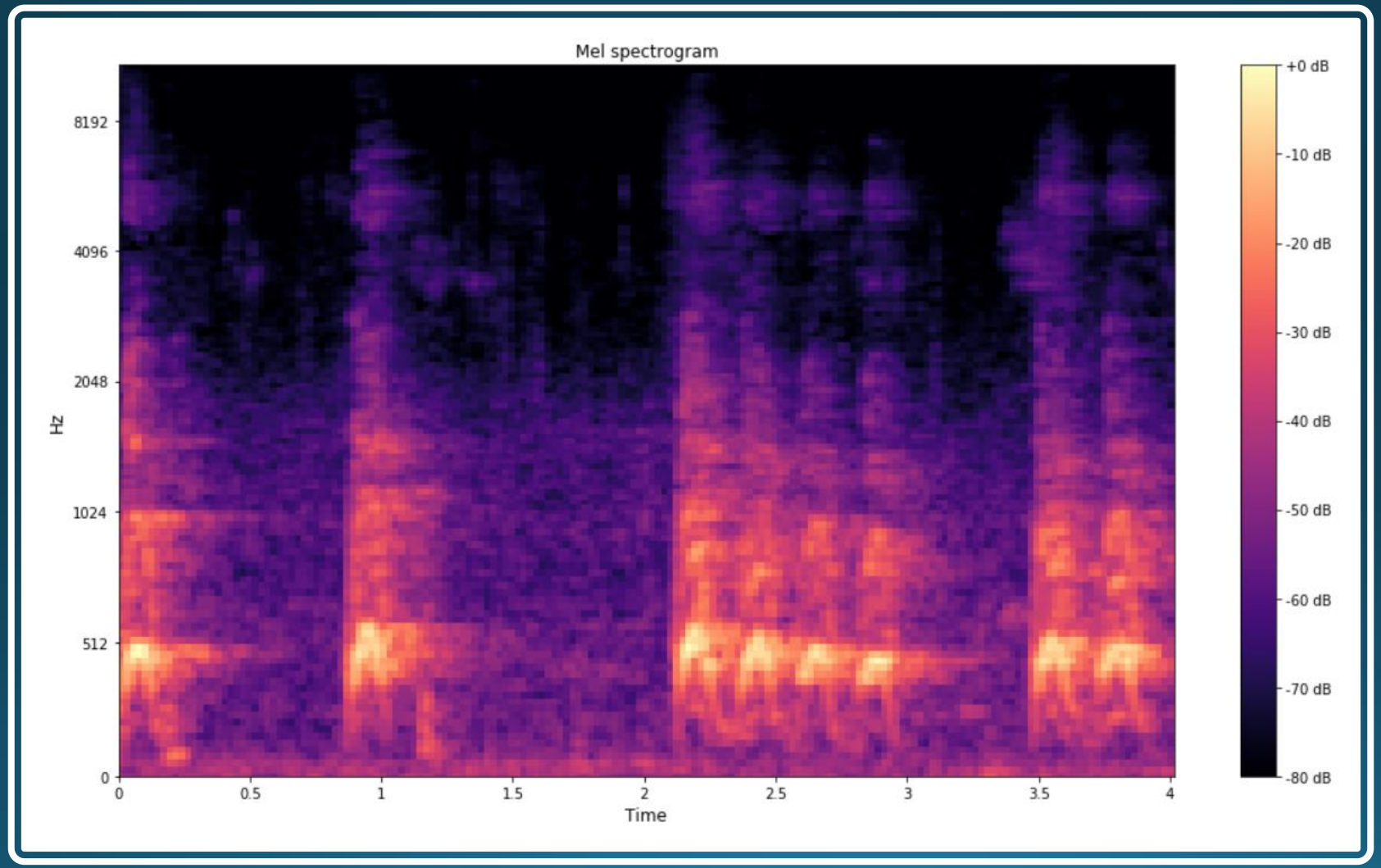


Enough Complaining!

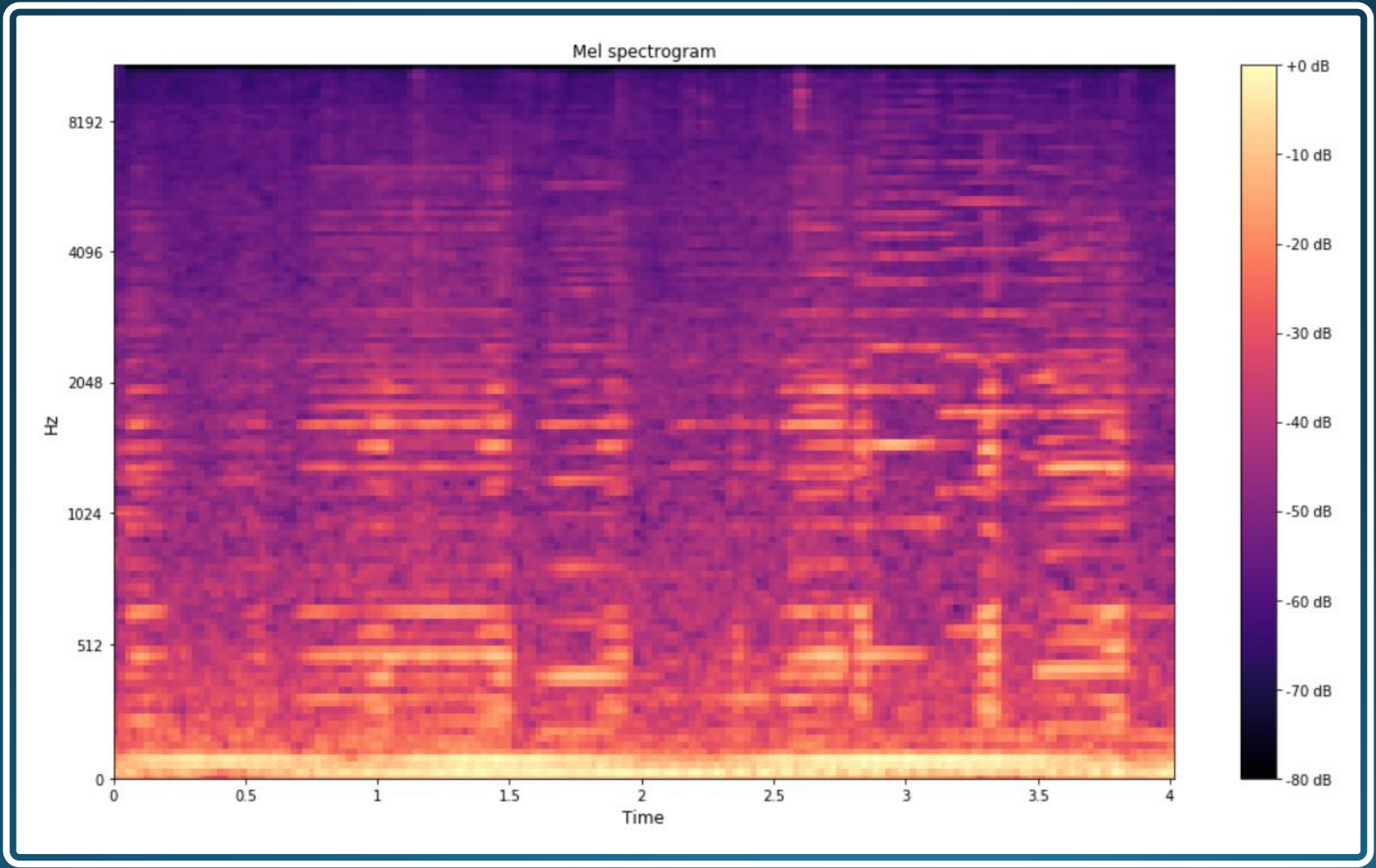
Here are the classes!

- Air Conditioner
- Car Horn
- Children Playing
- Dog Bark
- Drilling
- Engine Idling
- Gun Shot
- Jackhammer
- Siren
- Street Music

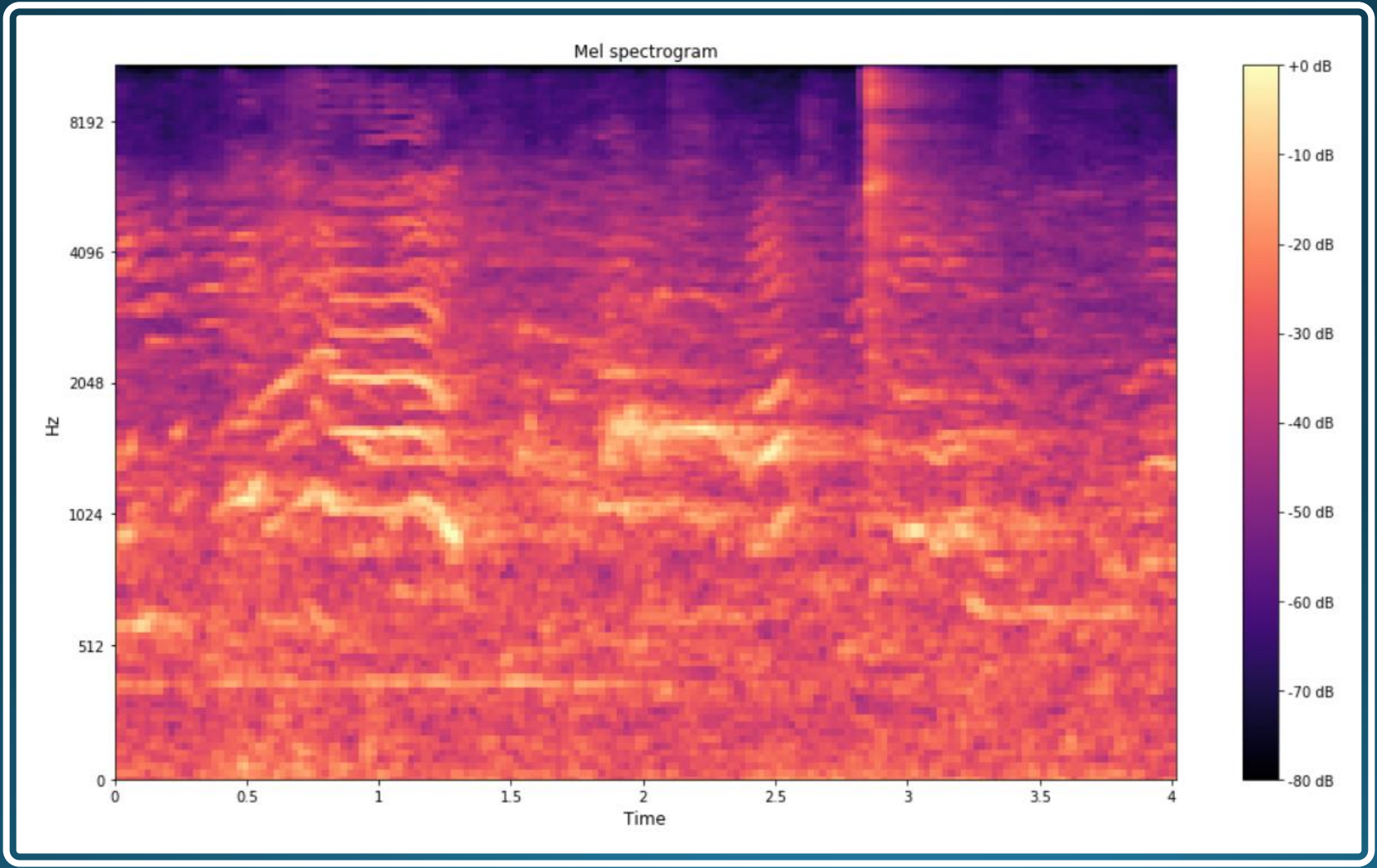
Dog Barking


Mel spectrogram
8192
4096
2048
1024
512
0
Hz
0
0.5
1
1.5
2
2.5
3
3.5
4
Time
+0 dB
-10 dB
-20 dB
-30 dB
-40 dB
-50 dB
-60 dB
-70 dB
-80 dB

Street Music

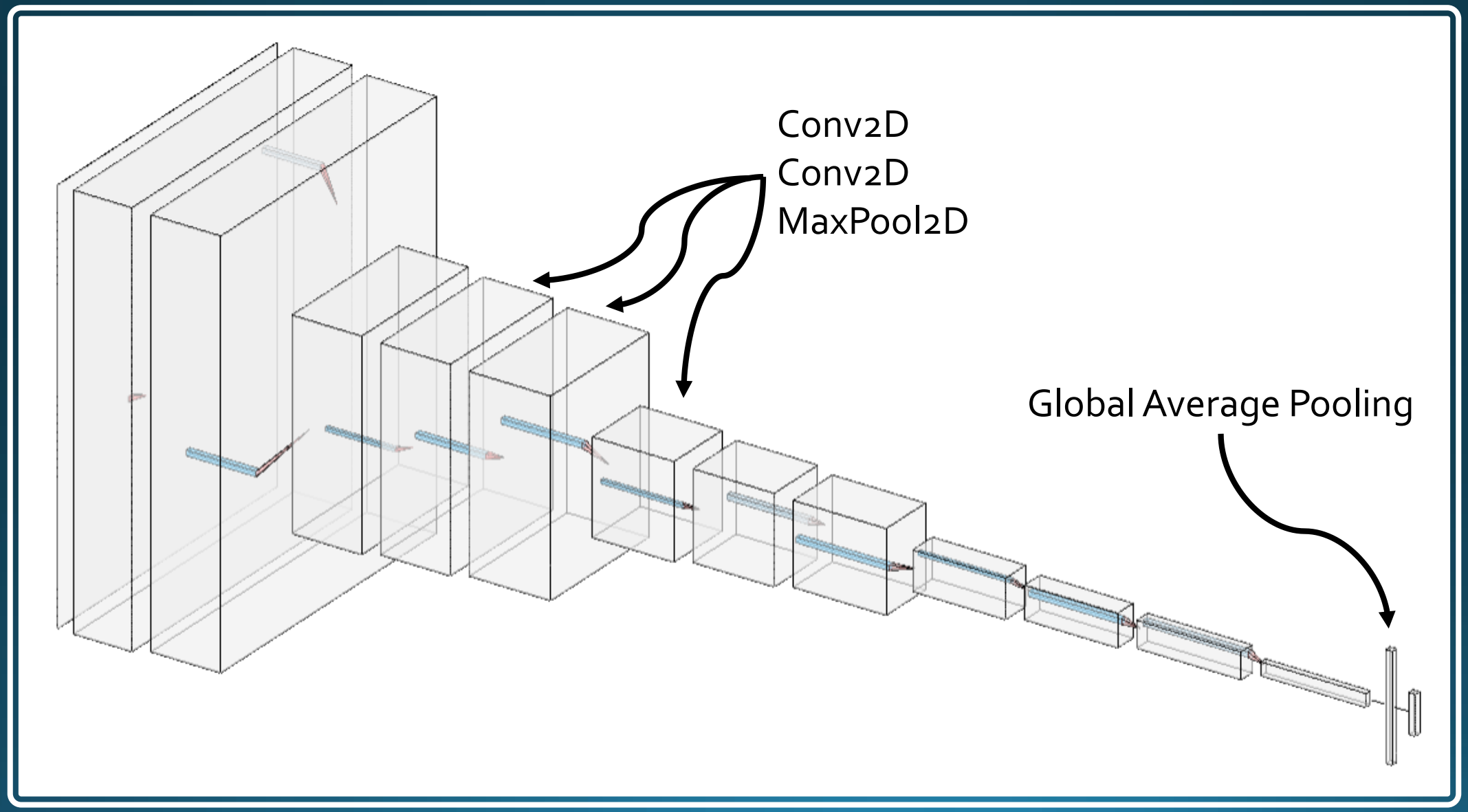


Children Playing

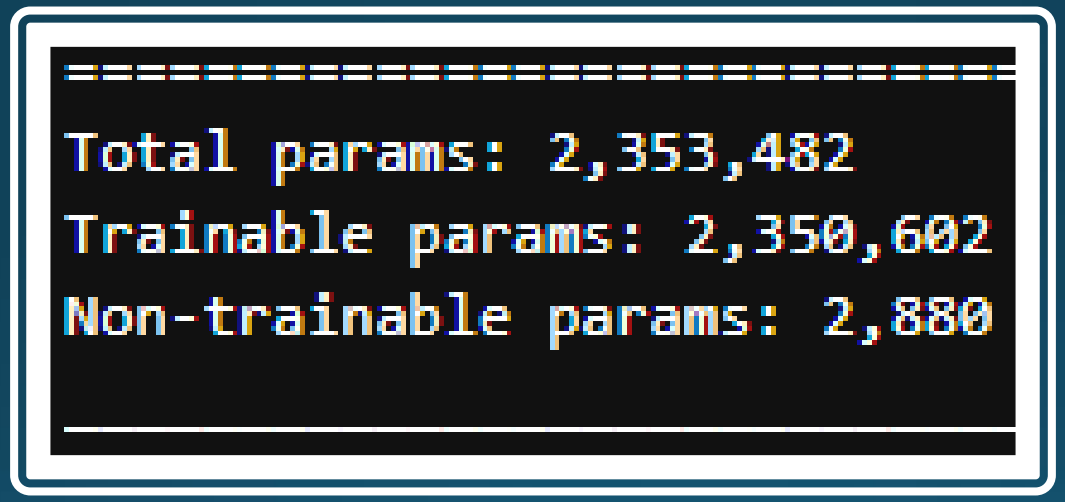


The hard part is over.
Now for Machine Learning.

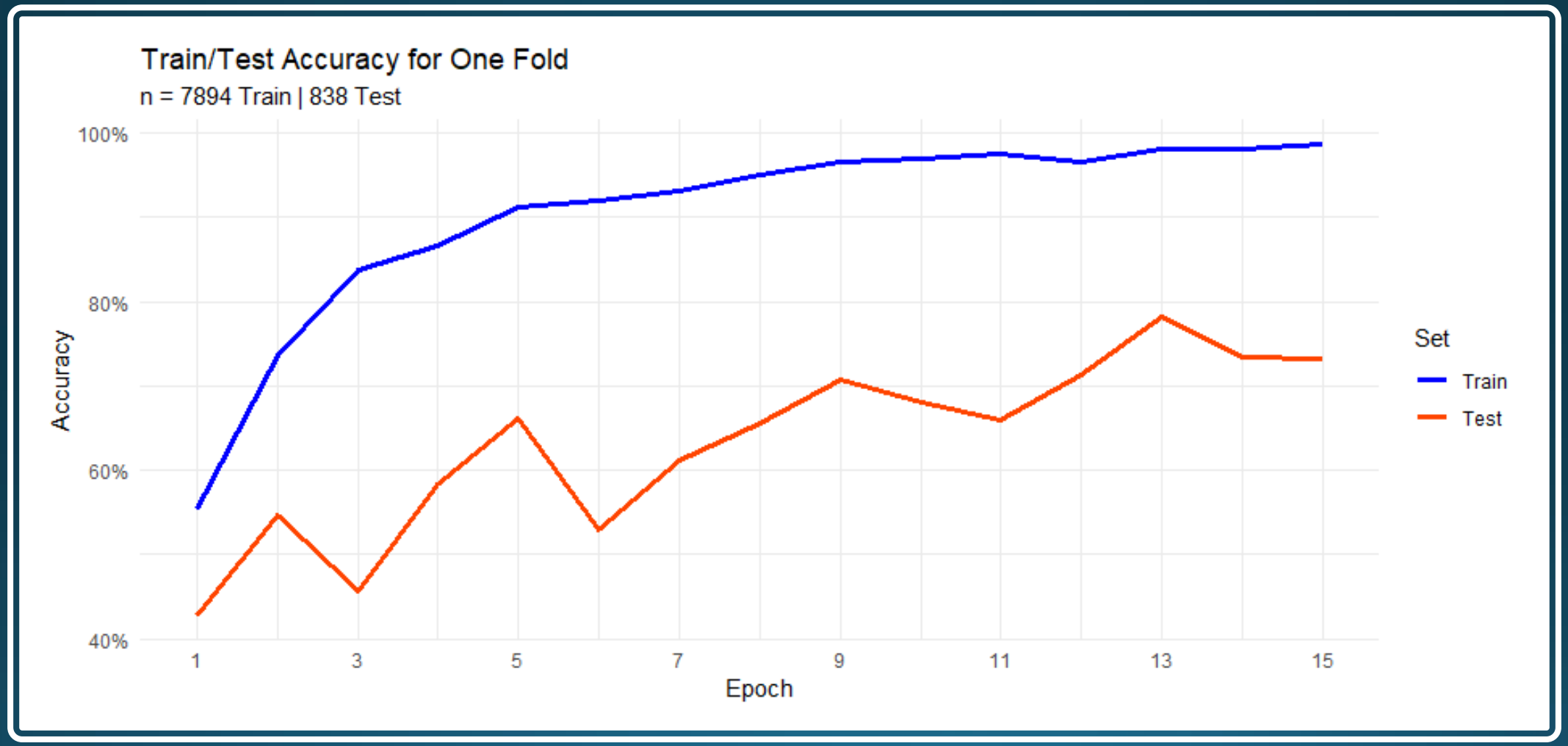
Baby's First Convolutional Neural Net

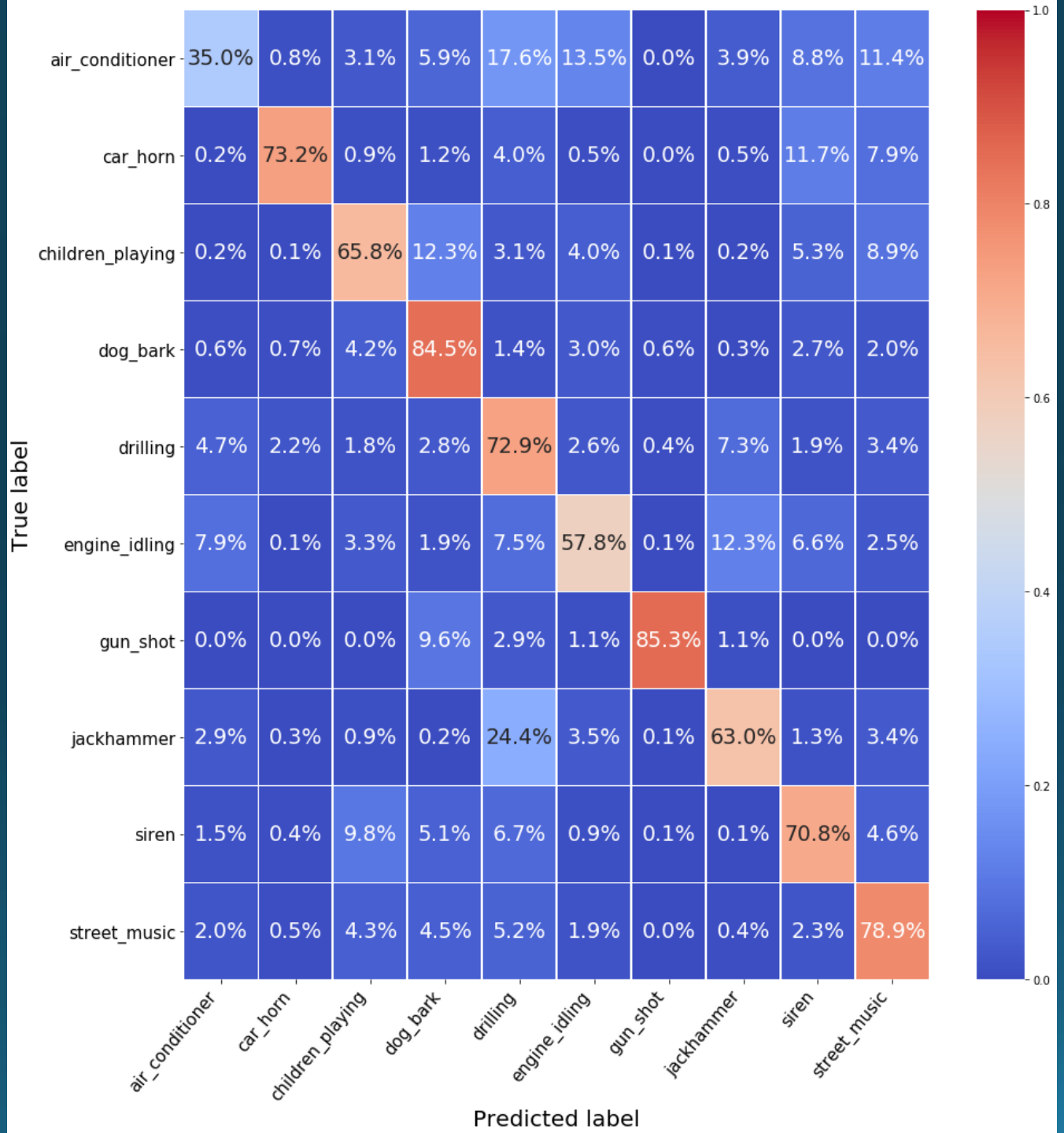


Very Small!


==========
Total params: 2,353,482
Trainable params: 2,350,602
Non-trainable params: 2,880
_____

Does it work?

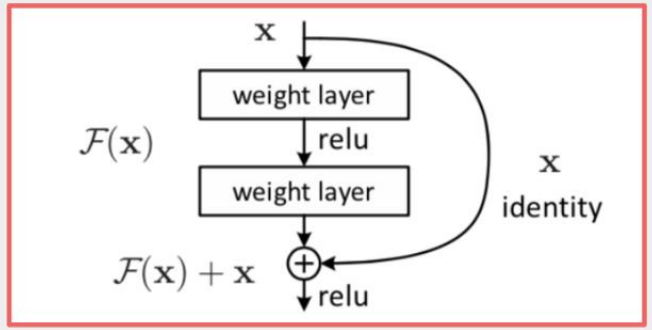

Train/Test Accuracy for One Fold
n = 7894 Train | 838 Test
100%
80%
60%
40%
Accuracy
1
3
5
7
9
11
13
15
Epoch
Set
Train
Test


air_conditioner 35.0% 0.8% 3.1% 5.9% 17.6% 13.5% 0.0% 3.9% 8.8% 11.4%
car_horn 0.2% 73.2% 0.9% 1.2% 4.0% 0.5% 0.0% 0.5% 11.7% 7.9%
children_playing 0.2% 0.1% 65.8% 12.3% 3.1% 4.0% 0.1% 0.2% 5.3% 8.9%
dog_bark 0.6% 0.7% 4.2% 84.5% 1.4% 3.0% 0.6% 0.3% 2.7% 2.0%
drilling 4.7% 2.2% 1.8% 2.8% 72.9% 2.6% 0.4% 7.3% 1.9% 3.4%
engine_idling 7.9% 0.1% 3.3% 1.9% 7.5% 57.8% 0.1% 12.3% 6.6% 2.5%
gun_shot 0.0% 0.0% 0.0% 9.6% 2.9% 1.1% 85.3% 1.1% 0.0% 0.0%
jackhammer 2.9% 0.3% 0.9% 0.2% 24.4% 3.5% 0.1% 63.0% 1.3% 3.4%
siren 1.5% 0.4% 9.8% 5.1% 6.7% 0.9% 0.1% 0.1% 70.8% 4.6%
street_music 2.0% 0.5% 4.3% 4.5% 5.2% 1.9% 0.0% 0.4% 2.3% 78.9%
True label
air_conditioner
car_horn
children_playing
dog_bark
drilling
engine_idling
gun_shot
jackhammer
siren
street_music
Predicted label
1.0
0.8
0.6
0.4
0.2
0.0

67%

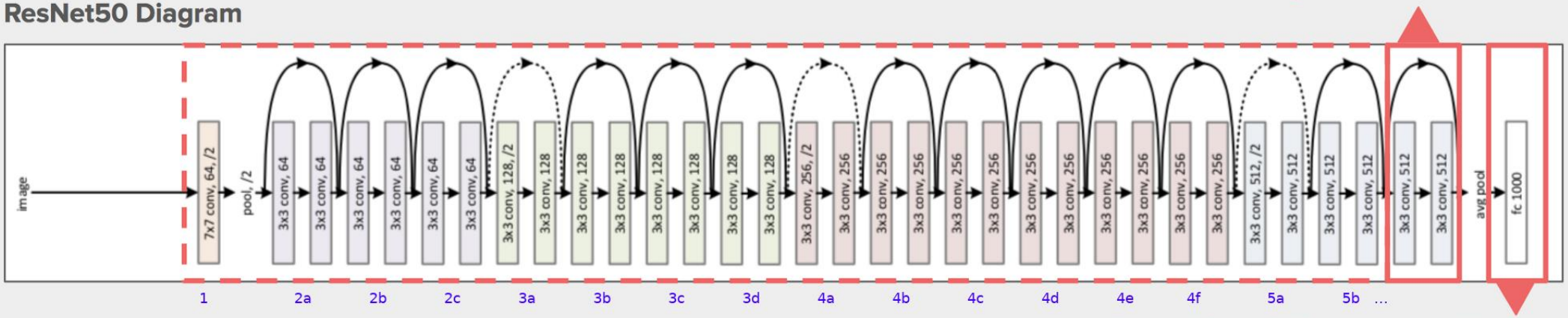
Okay, time for the big guns.

Retrain ResNet50

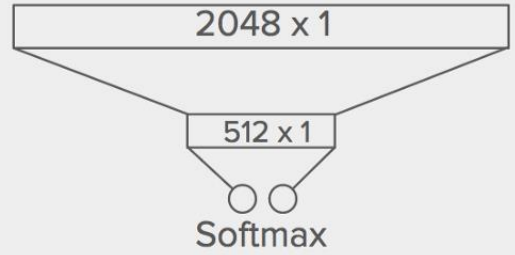


Residual Learning Block

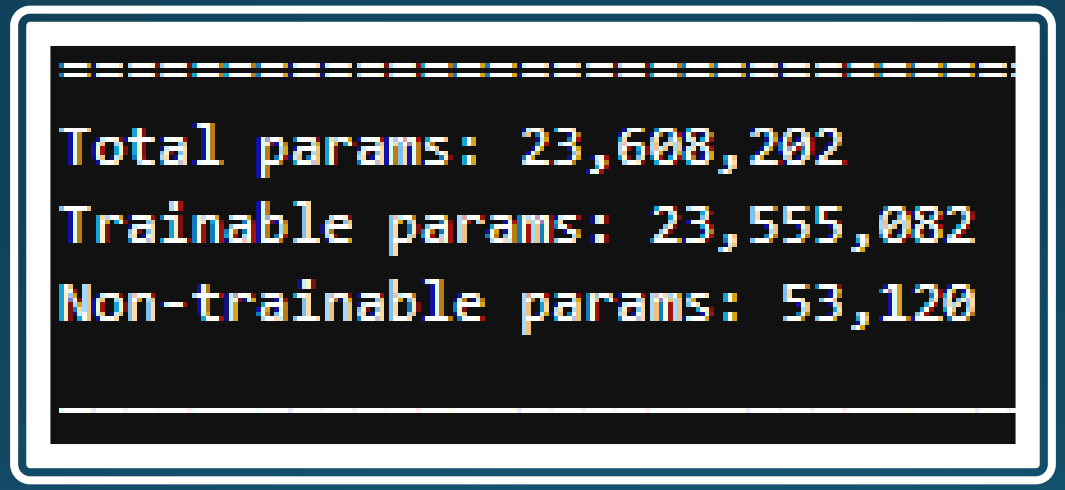
ResNet50 Diagram

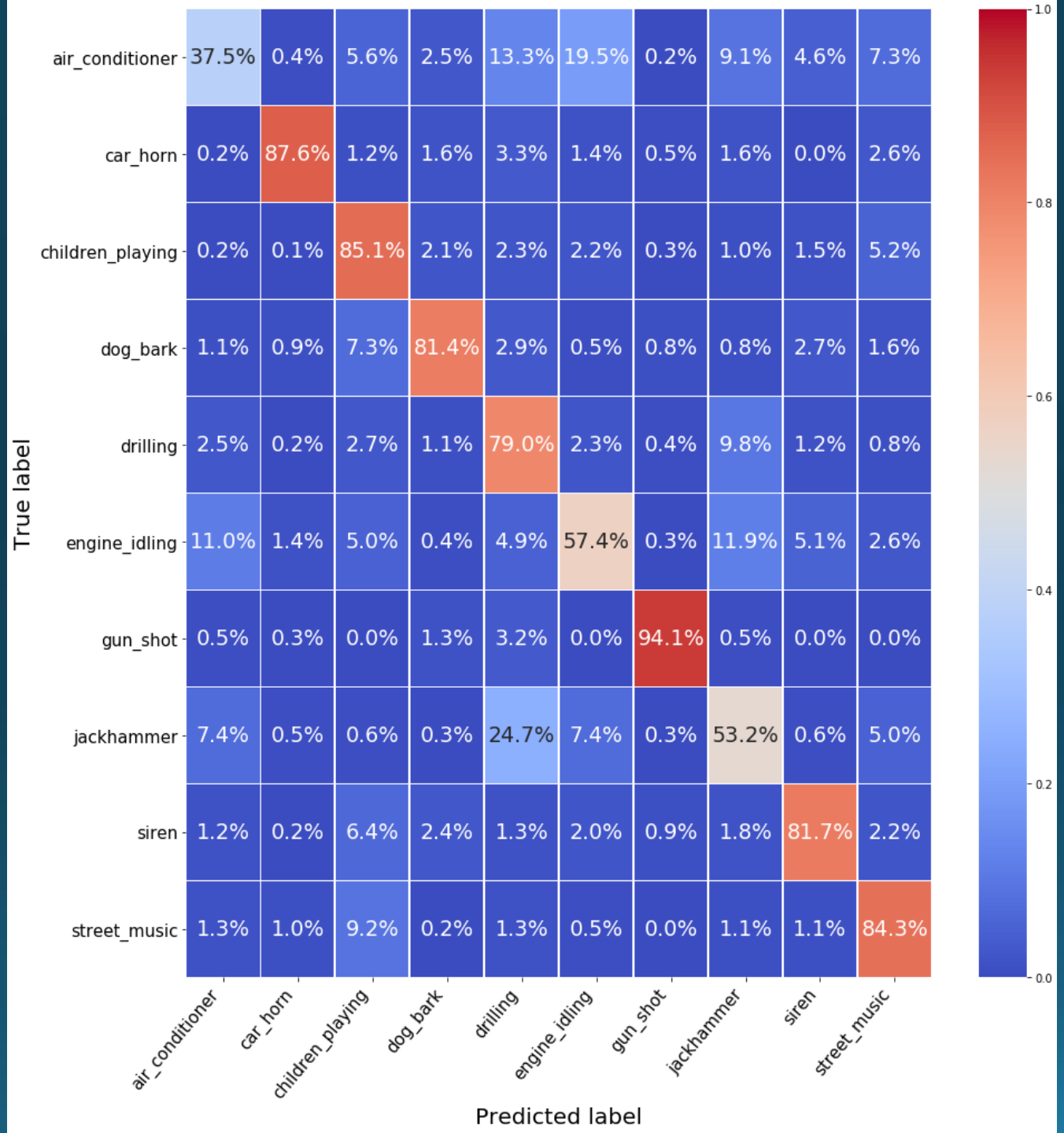


Re-architect fully-connected layers



Not Very Small...




air_conditioner 37.5% 0.4% 5.6% 2.5% 13.3% 19.5% 0.2% 9.1% 4.6% 7.3%
car_horn 0.2% 87.6% 1.2% 1.6% 3.3% 1.4% 0.5% 1.6% 0.0% 2.6%
children_playing 0.2% 0.1% 85.1% 2.1% 2.3% 2.2% 0.3% 1.0% 1.5% 5.2%
dog_bark 1.1% 0.9% 7.3% 81.4% 2.9% 0.5% 0.8% 0.8% 2.7% 1.6%
drilling 2.5% 0.2% 2.7% 1.1% 79.0% 2.3% 0.4% 9.8% 1.2% 0.8%
engine_idling 11.0% 1.4% 5.0% 0.4% 4.9% 57.4% 0.3% 11.9% 5.1% 2.6%
gun_shot 0.5% 0.3% 0.0% 1.3% 3.2% 0.0% 94.1% 0.5% 0.0% 0.0%
jackhammer 7.4% 0.5% 0.6% 0.3% 24.7% 7.4% 0.3% 53.2% 0.6% 5.0%
siren 1.2% 0.2% 6.4% 2.4% 1.3% 2.0% 0.9% 1.8% 81.7% 2.2%
street_music 1.3% 1.0% 9.2% 0.2% 1.3% 0.5% 0.0% 1.1% 1.1% 84.3%
True label
air_conditioner
car_horn
children_playing
dog_bark
drilling
engine_idling
gun_shot
jackhammer
siren
street_music
Predicted label
1.0
0.8
0.6
0.4
0.2
0.0

72%

We did it!