Speech Recognition with TensorFlow

Vira T Capell

Applied Data Science — Data Stories

November 6 2019

Machine Learning
Speech Recognition Usage

About a dataset

- 917 wav files,
- 13 spoken numbers,
- many different people spoken:

  both genders,

  different ascents

00_66a1550b08.wav

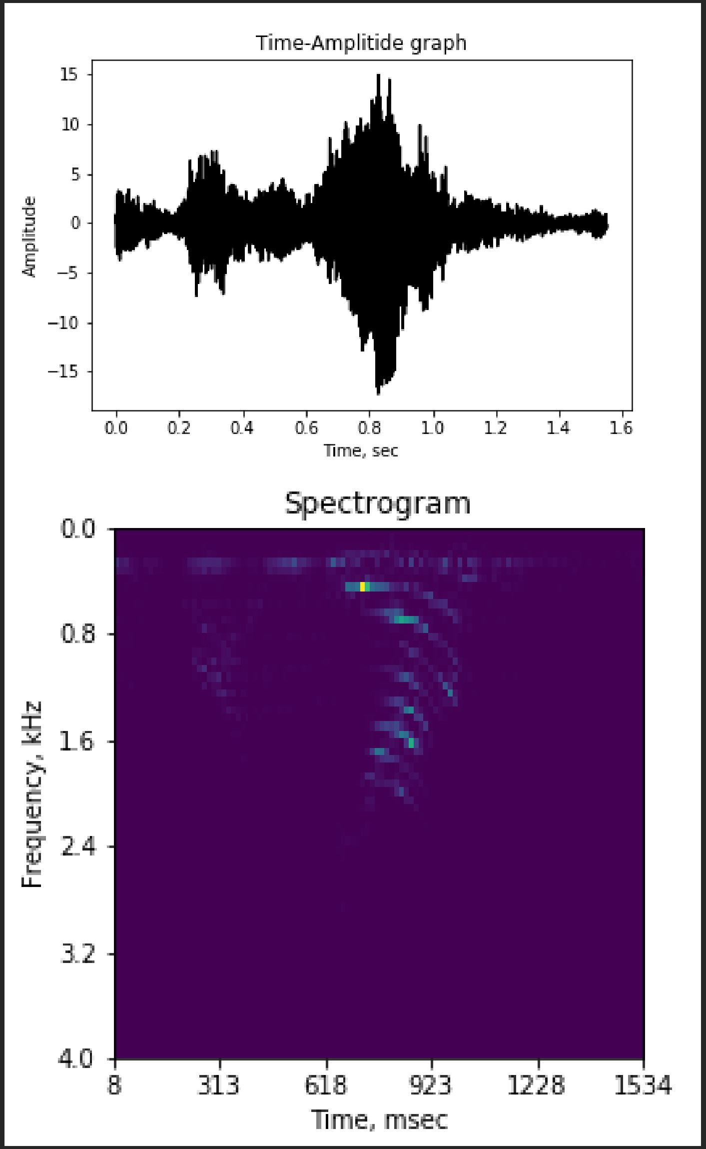
Sample Rate=8 kHz

Usable voice frequency band:
0.3 to 3.4 kHz

SciPy.org Sponsored By ENTHOUGHT

```
import scipy.io.wavfile as wv

SampleRate,data = wv.read('filename.wav')
```

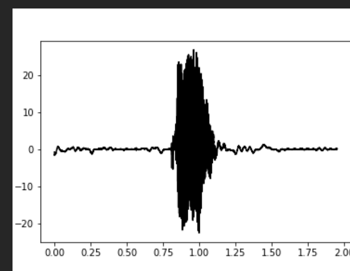
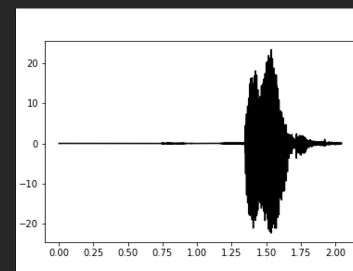
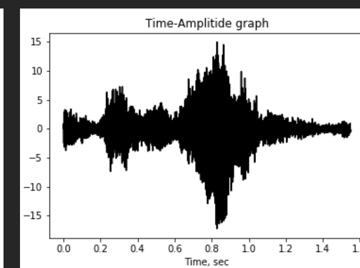


Team

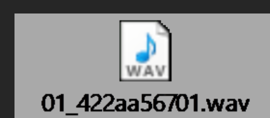
- Allison Wong
- Andy Houseman
- Michelle Duer
- Tal Stoner
- Tom Widdows
- Vira T Capell

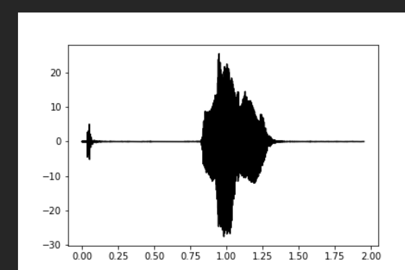
Eyeballing data

- "Silence" before and after a number

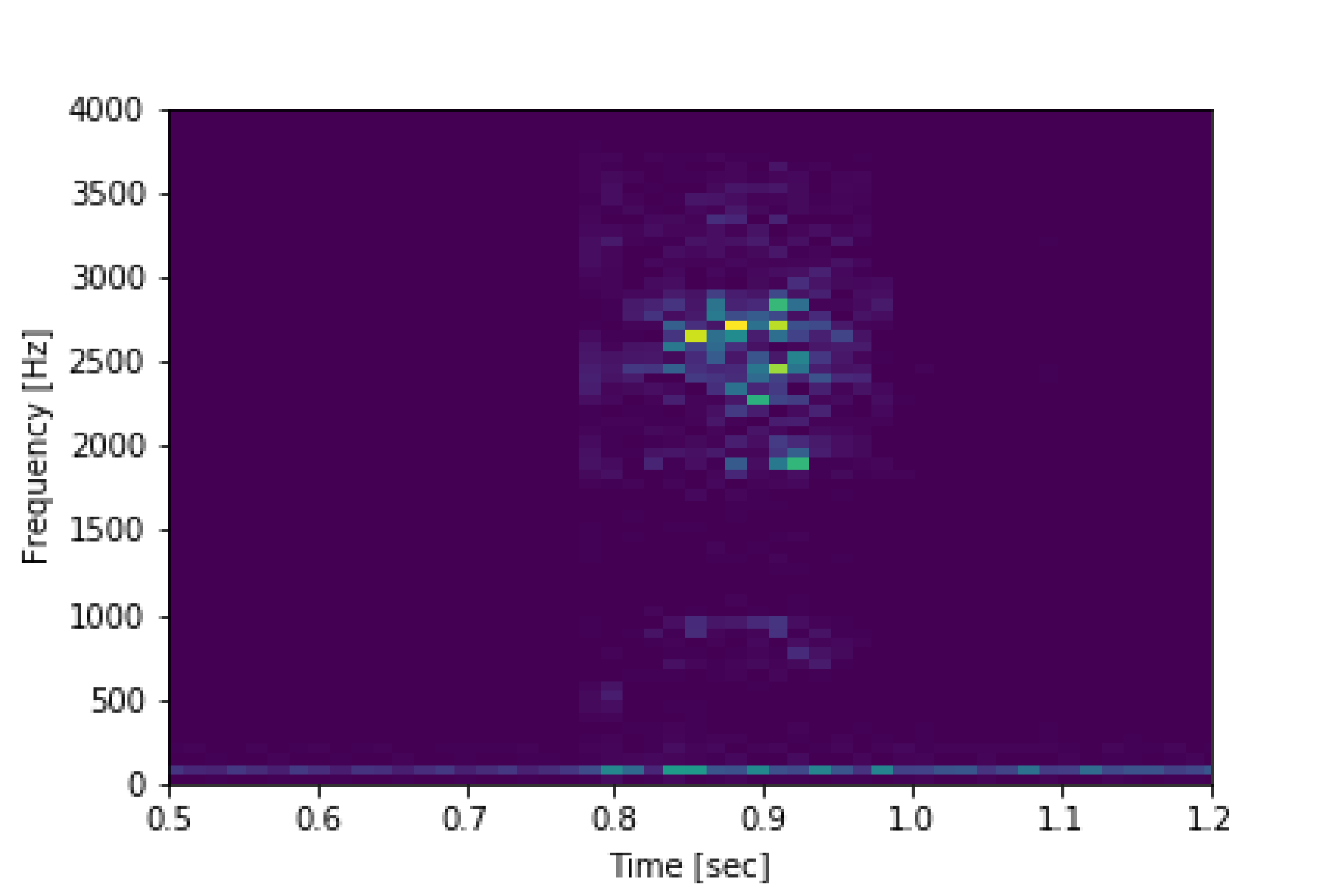




- Noisy: background noise, "click" sound on beginning



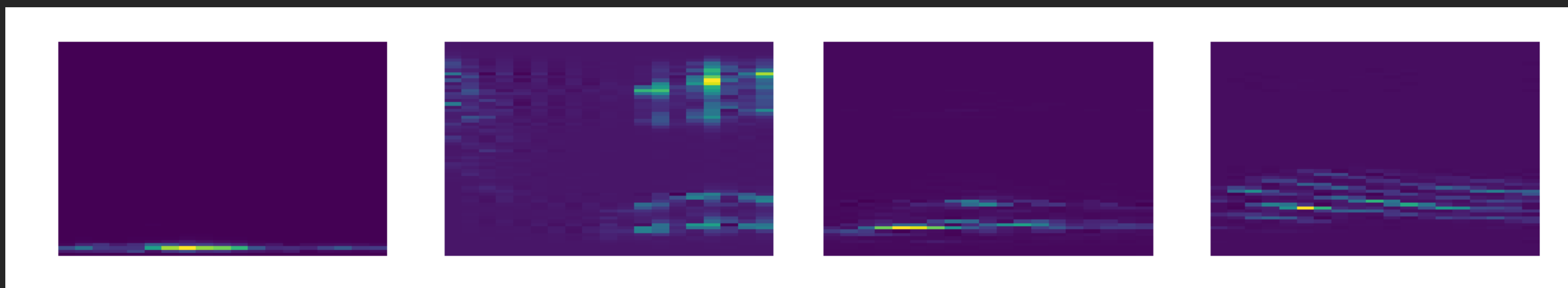


Spectrogram

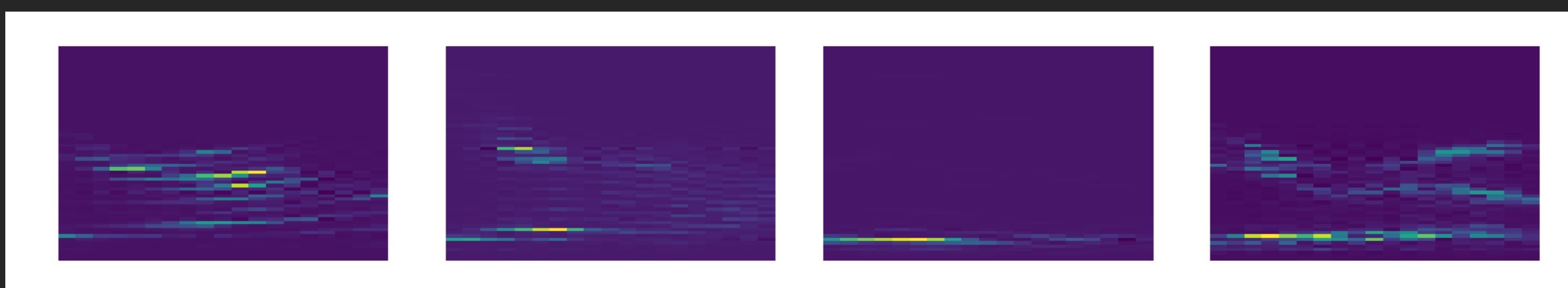


Look different for same number spoken by different people

"12"

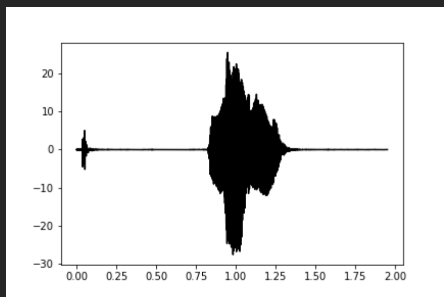


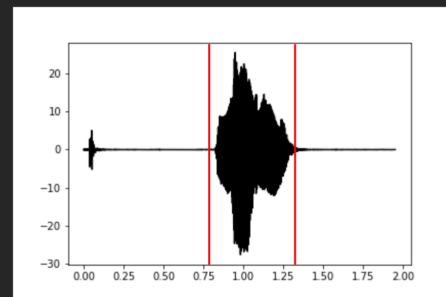
"0"

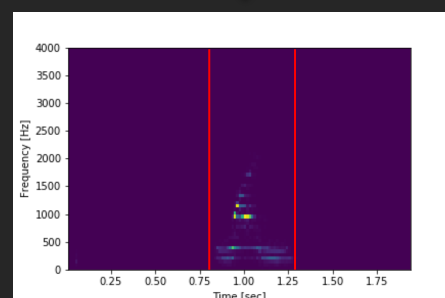


Cleaning the data

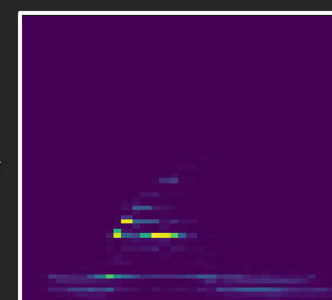
WAV file







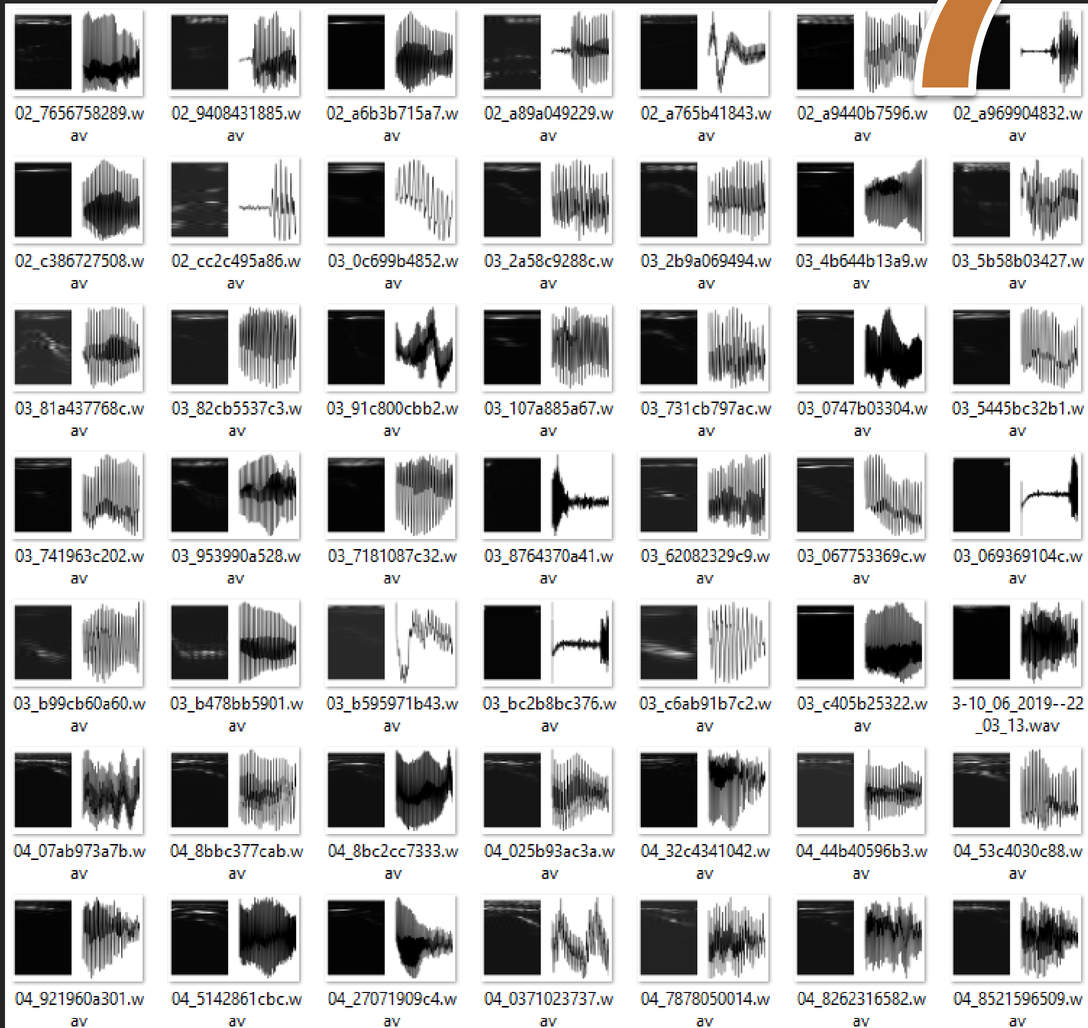
PNG file



Idea: using TensorFlow to train the CNN to recognize spectrograms as pictures

Michelle Duer:
Accuracy 60%

Both spectrograms and amplitude traces to TensorFlow



Model

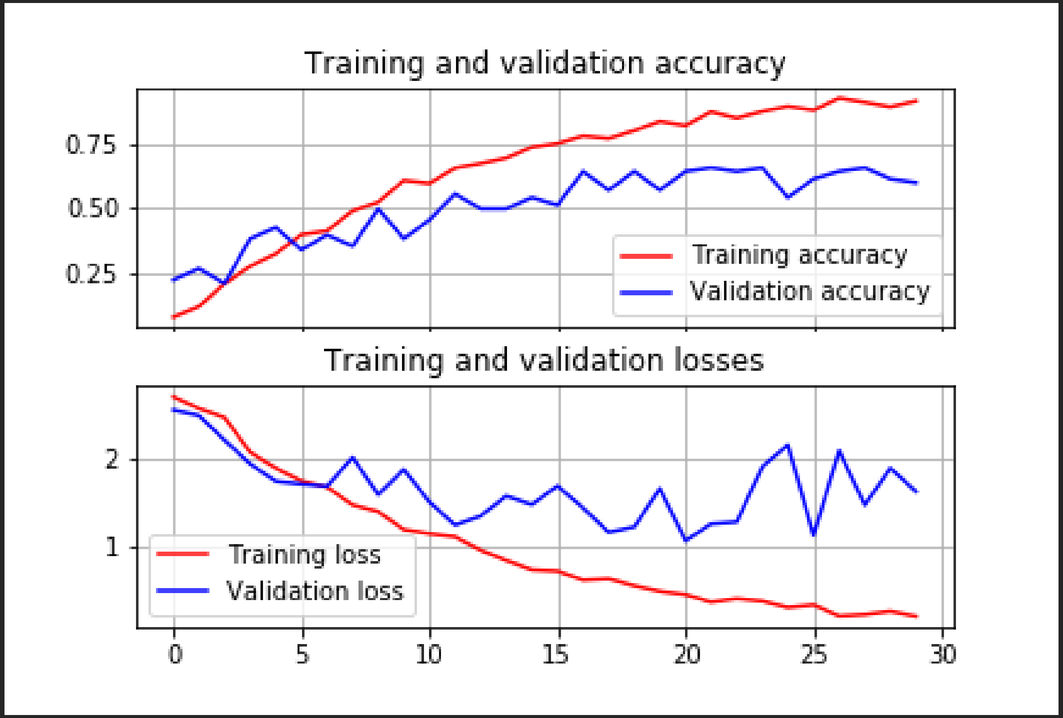
Layer (type)	Output Shape	Param #
conv2d_30 (Conv2D)	(None, 198, 198, 140)	3920
max_pooling2d_30 (MaxPooling	(None, 99, 99, 140)	0
conv2d_31 (Conv2D)	(None, 98, 98, 64)	35904
max_pooling2d_31 (MaxPooling	(None, 49, 49, 64)	0
conv2d_32 (Conv2D)	(None, 47, 47, 128)	73856
max_pooling2d_32 (MaxPooling	(None, 23, 23, 128)	0
conv2d_33 (Conv2D)	(None, 21, 21, 128)	147584
max_pooling2d_33 (MaxPooling	(None, 10, 10, 128)	0
flatten_8 (Flatten)	(None, 12800)	0
dropout_8 (Dropout)	(None, 12800)	0
dense_16 (Dense)	(None, 512)	6554112
dense_17 (Dense)	(None, 13)	6669

Total params: 6,822,045
Trainable params: 6,822,045
Non-trainable params: 0

Performance: over 65% on validation set

val_acc[20]
0.6571429

After about 20 epochs validation accuracy reaches %65, and climbs further up, but the validation loss is going up after 20 epochs, showing over fitting of the network



Performance on training set

X = wav_to_png (filename.wav)
prediction = model.predict(X)

File name	Spoken Number	Predicted
006a5b2cb6.wav	11	2
0130bbb543.wav	4	4
0217a3b12a.wav	2	2
029830a064.wav	6	6
099743a386.wav	8	2
0c7156ca20.wav	2	2
135451ab23.wav	2	2
1510c30707.wav	0	1
16ac11144b.wav	4	11
17b6a64673.wav	7	0
1aa7222c02.wav	0	0
2581723976.wav	7	7
2885430727.wav	2	6
2918750a33.wav	3	2
298a616a81.wav	8	2
2aa949c3a3.wav	10	11
2b9c815b22.wav	2	1
2c2caacb08.wav	0	7
2ccc498b11.wav	1	1

Acknowledgments

Team

- Tom Widdows
- Tal Stoner
- Michelle Duer
- Andy Houseman
- Allison Wong
- Vira T Capell

Data Stories Hosts

- John Burt
- Mathew A. Borthwick, Ph.D.

Consulting:

Fritz Capell

Questions?

Fourrier Transformation