Kaggle Speech Recognition Challenge

Сиганов Илья, разработчик 7bits, ОмГУ

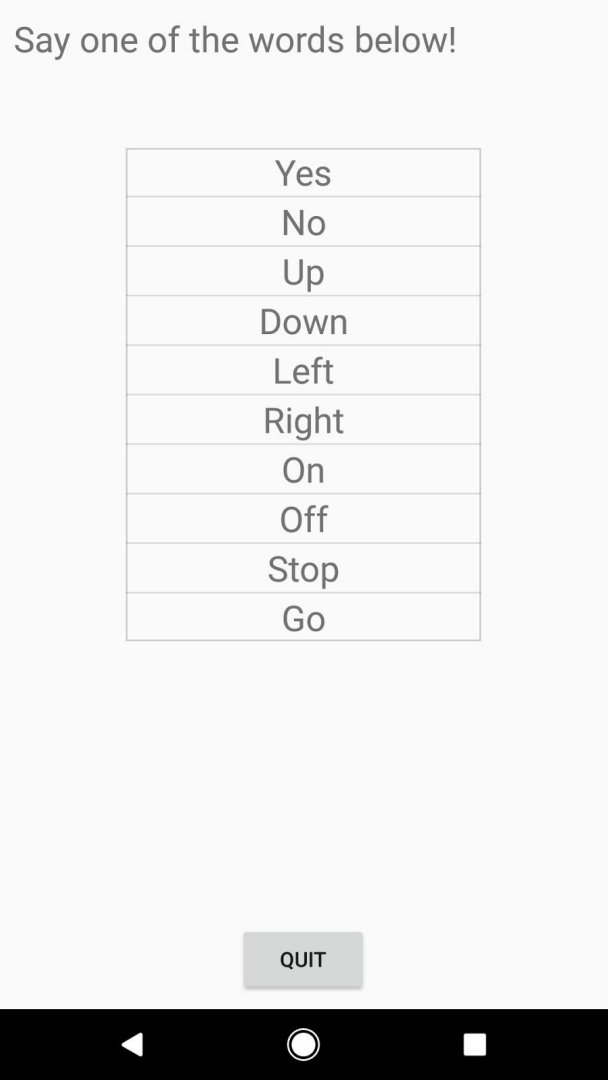
Организаторы

Google brain

Kaggle

Google cloud platform $500

Призовой фонд: $25,000



Задача

1 секундный клипы 16Кгц

12 классов

модель < 5 Мб

выполнение < 200мс на RPi

Данные

Train: 57929

Validation: 6798

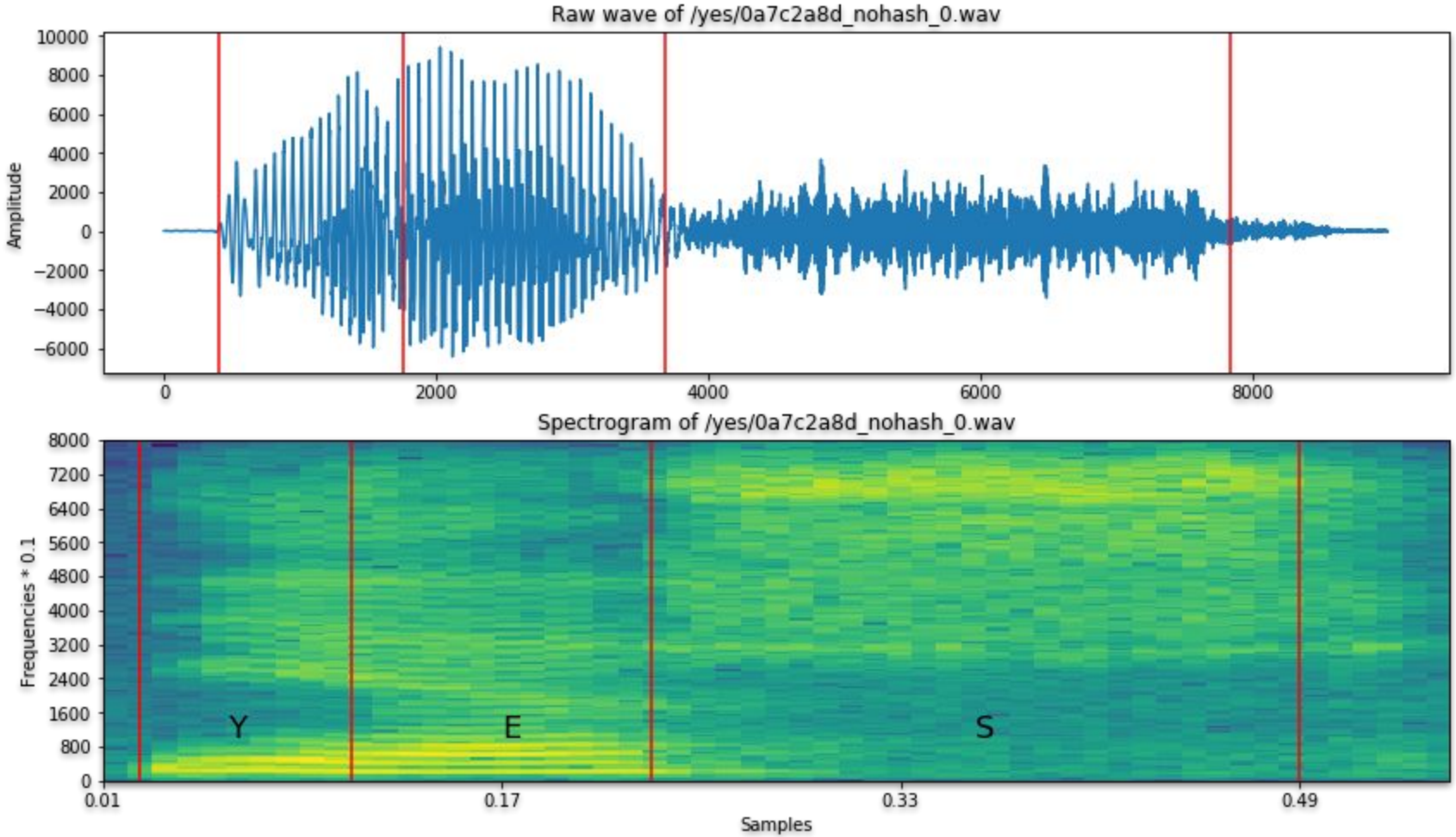
Test: 158538

Users: 1698

label	
silence	6
down	2095
off	2101
no	2105
left	2106
on	2110
right	2111
go	2112
up	2115
yes	2116
stop	2134
unknown	36818

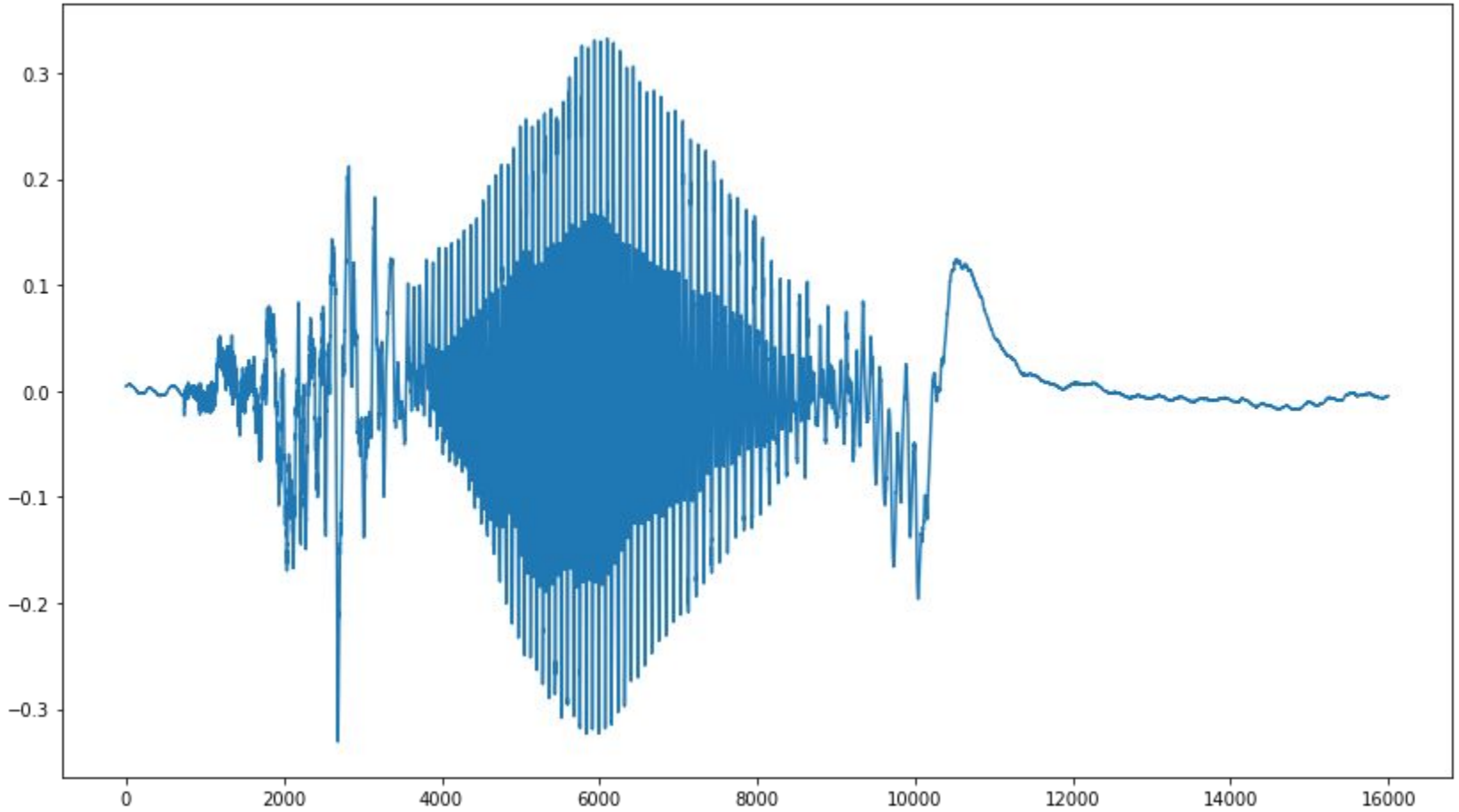
label	
stop	246
left	247
off	256
right	256
on	257
go	260
up	260
yes	261
down	264
no	270
unknown	4221

Как выглядит осциллограмма?

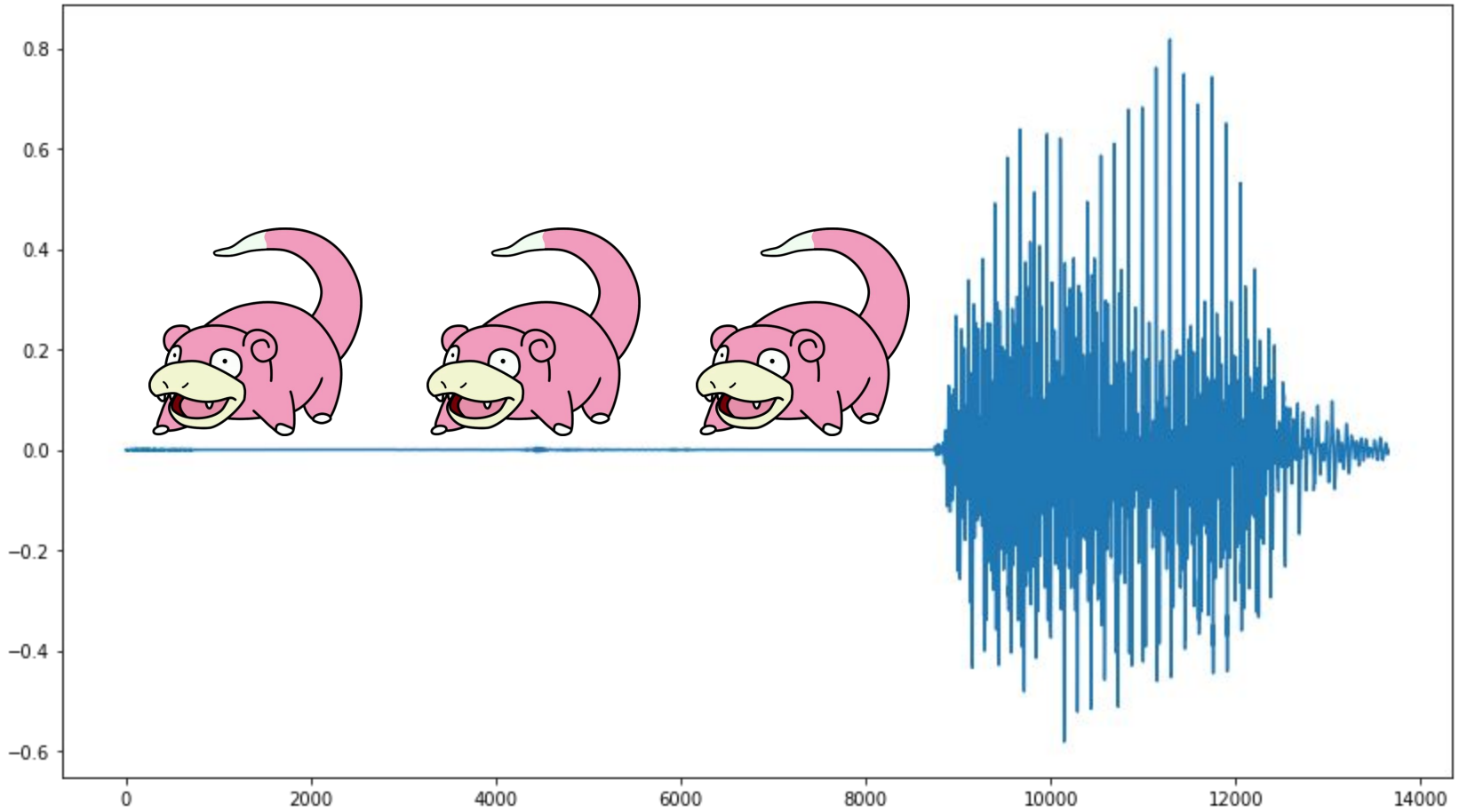


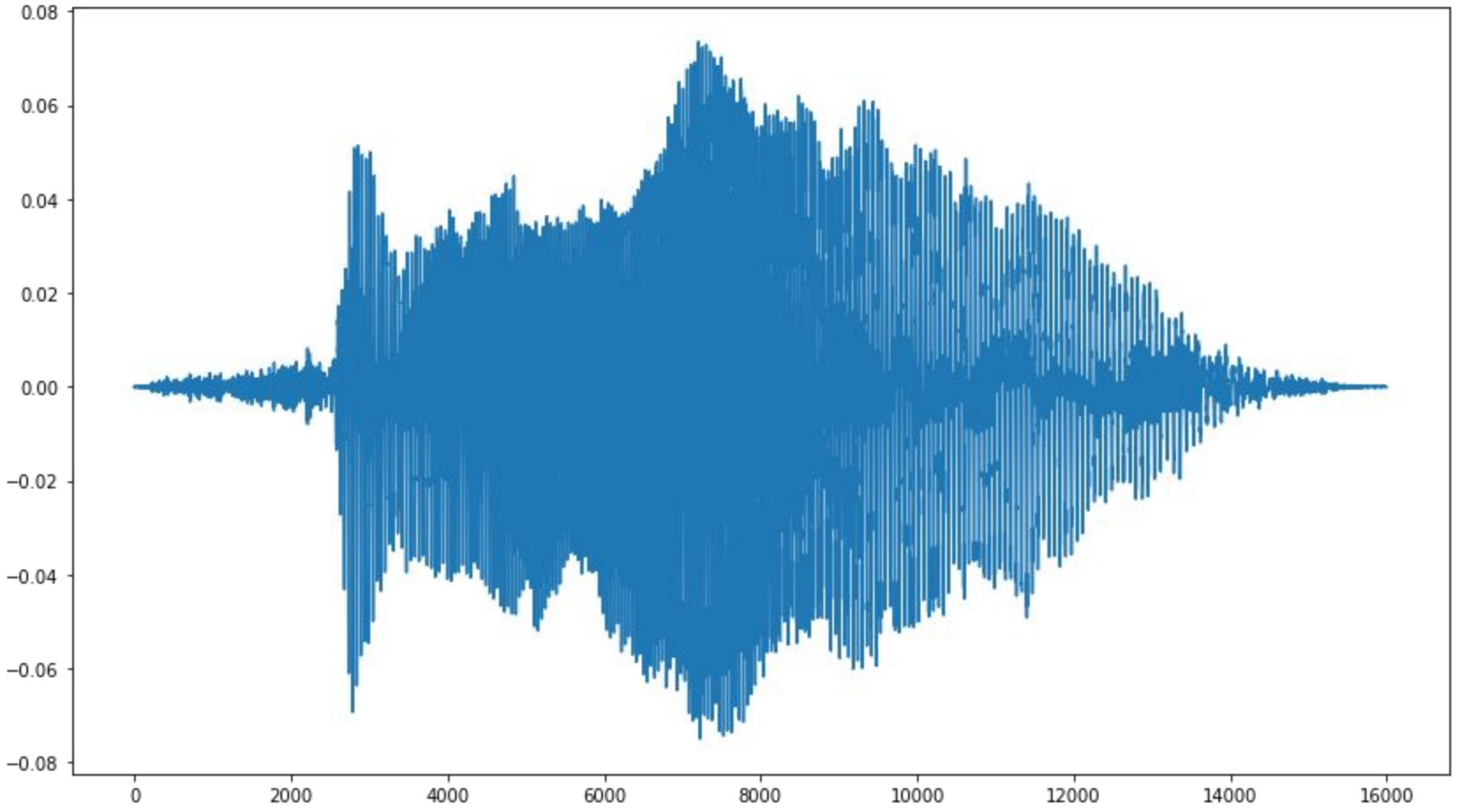
https://www.kaggle.com/davids1992/speech-representation-and-data-exploration

Как на самом деле выглядит осциллограмма

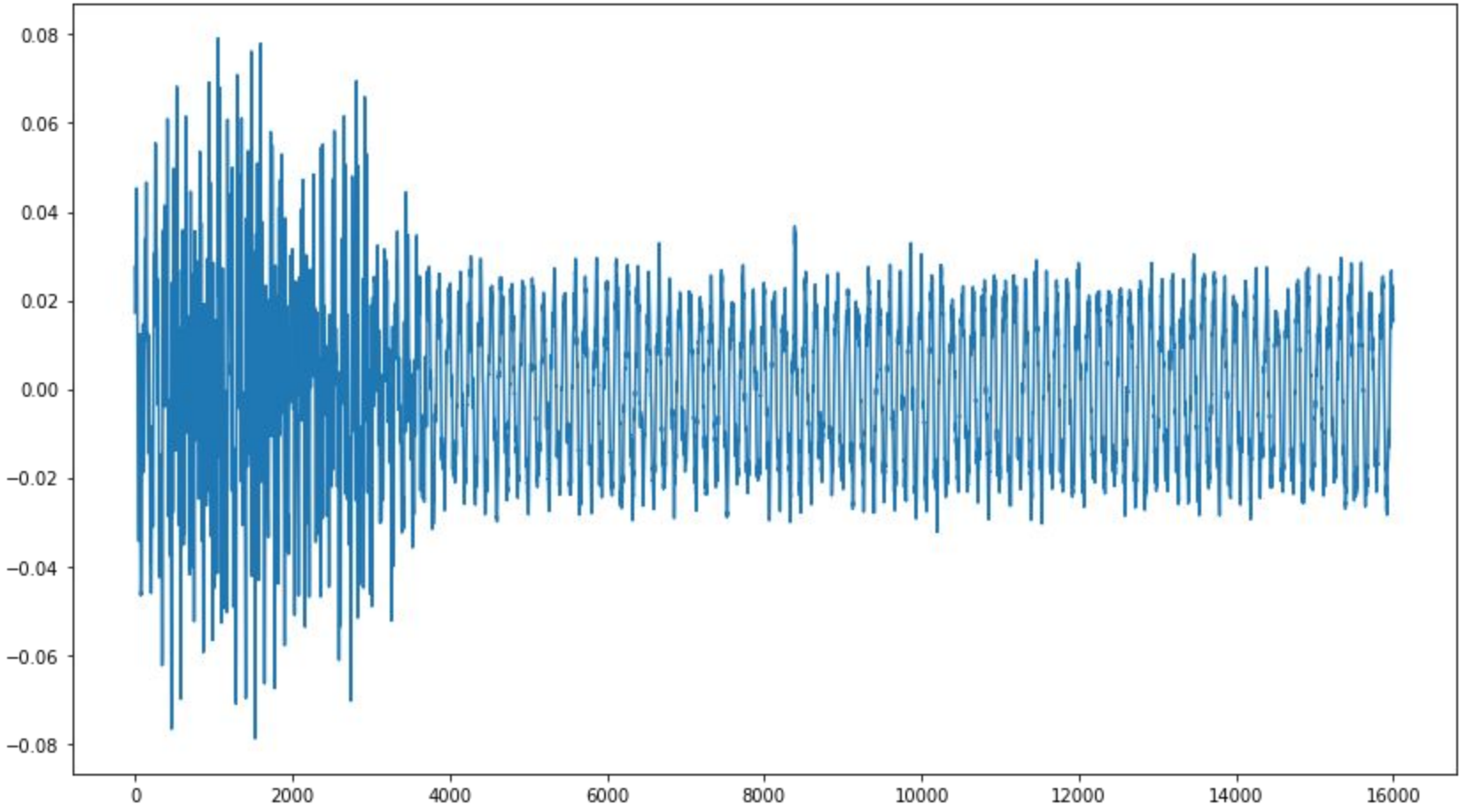

0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
2000
4000
6000
8000
10000
12000
14000
16000

tree


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
2000
4000
6000
8000
10000
12000
14000

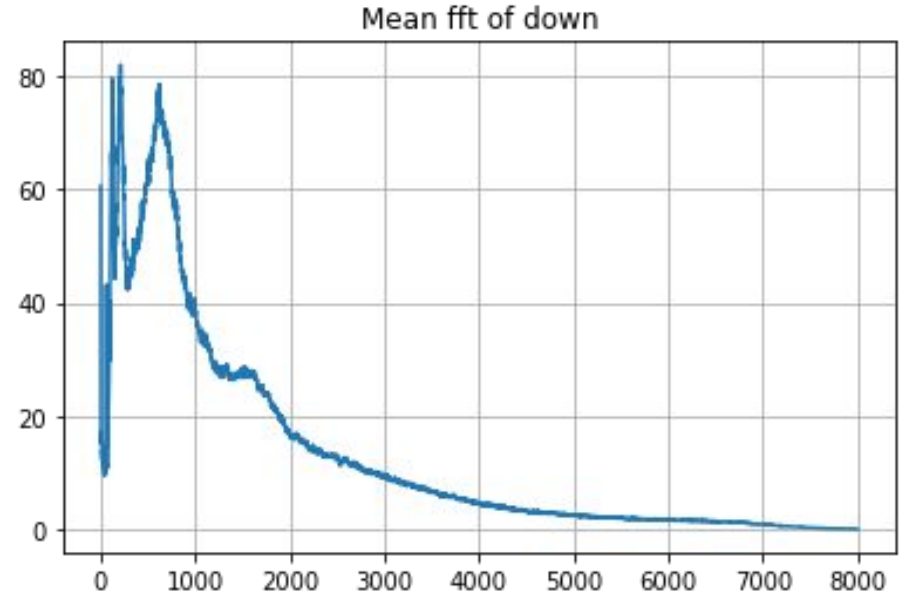

0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
0
2000
4000
6000
8000
10000
12000
14000
16000

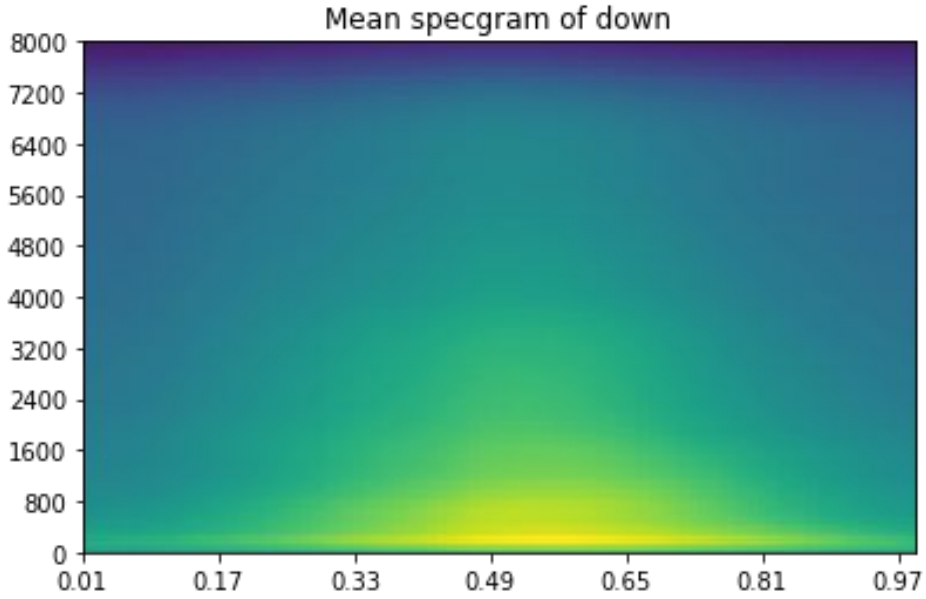
three


0.08
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−0.08
0
2000
4000
6000
8000
10000
12000
14000
16000

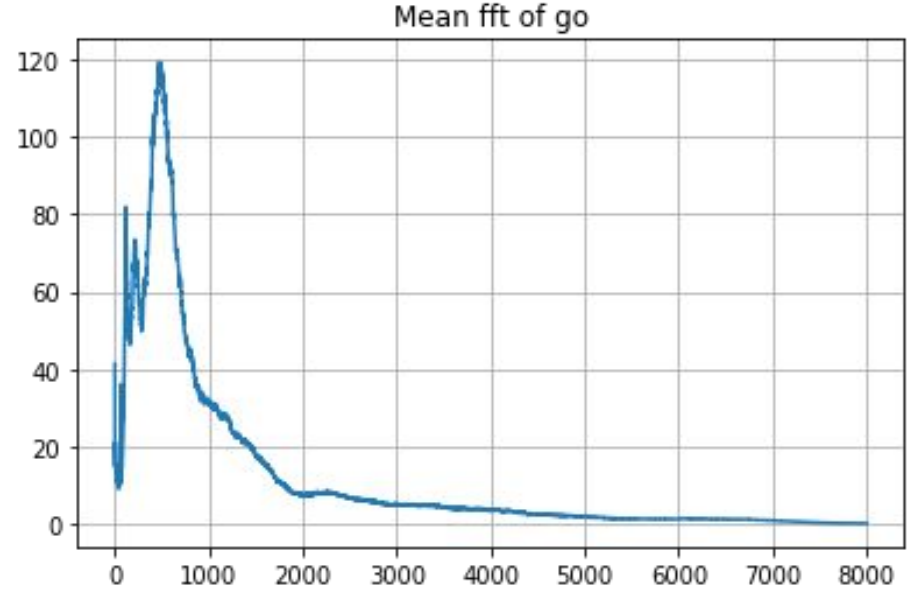
four

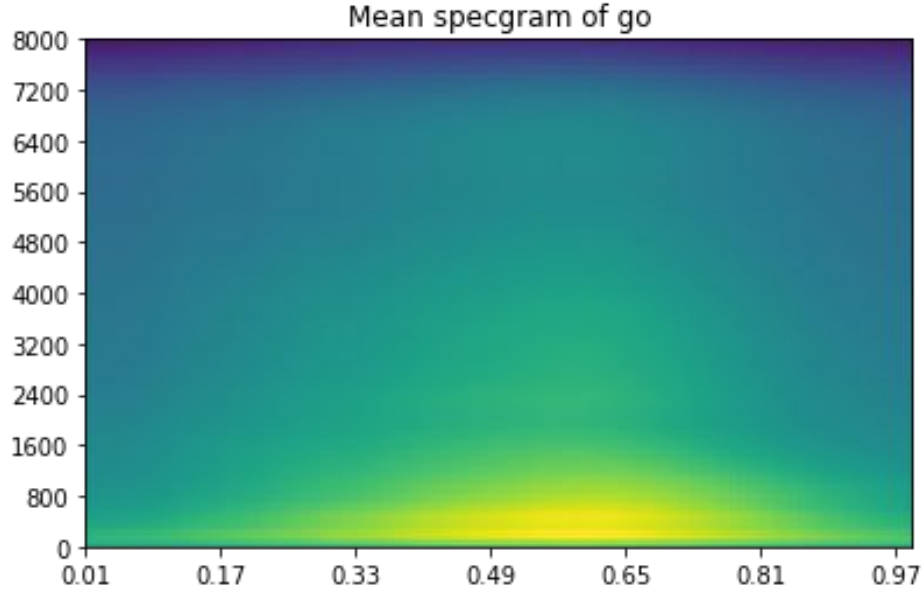
Первичный анализ



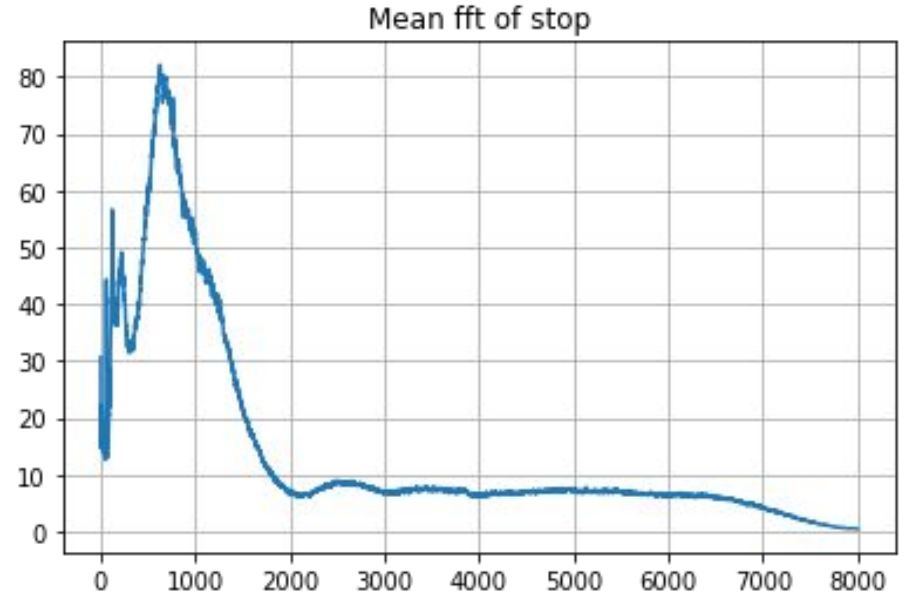


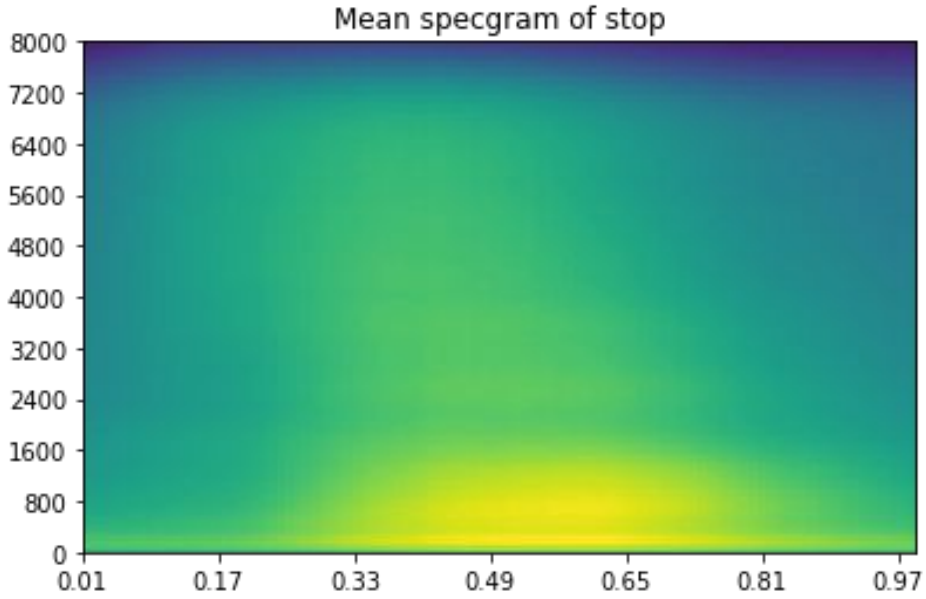
https://www.kaggle.com/davids1992/speech-representation-and-data-exploration





https://www.kaggle.com/davids1992/speech-representation-and-data-exploration



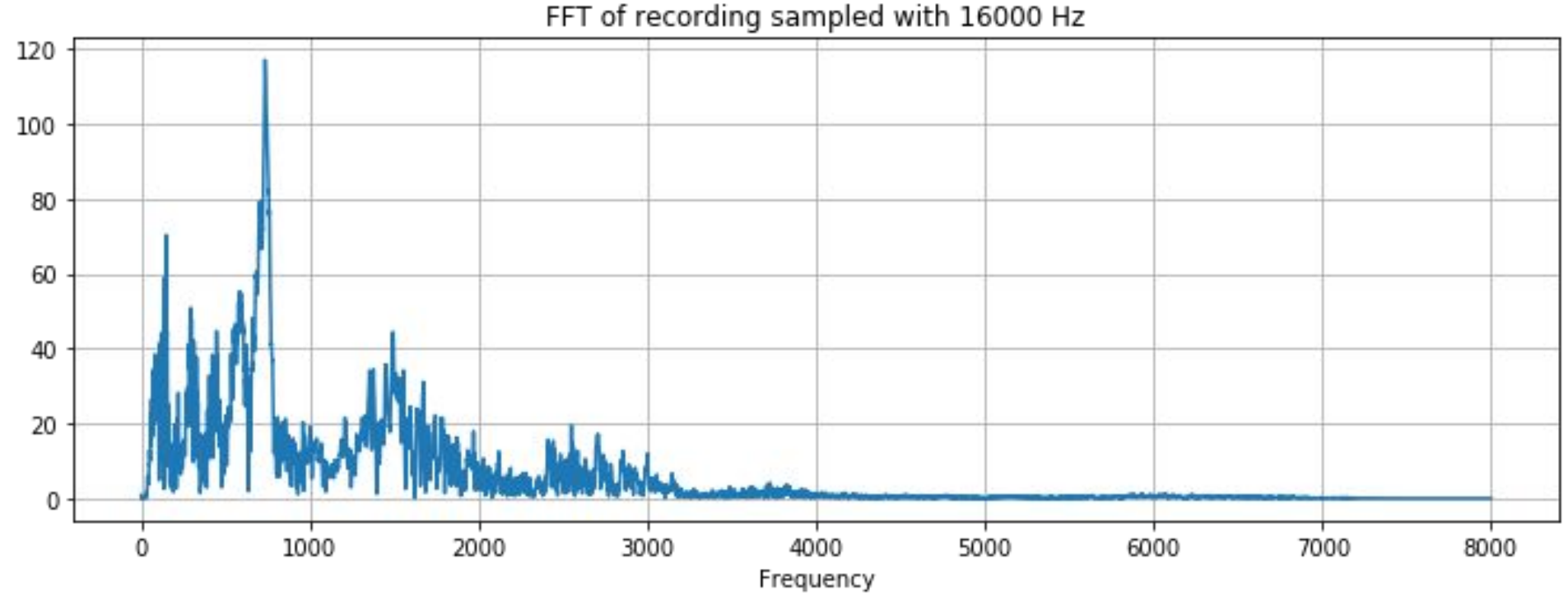


https://www.kaggle.com/davids1992/speech-representation-and-data-exploration

Признаки

FFT - переход в частотный домен

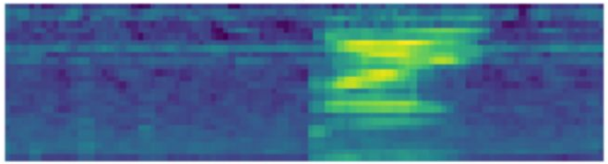
Спектральный анализ



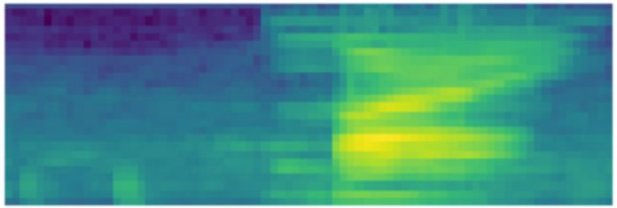
https://www.kaggle.com/davids1992/speech-representation-and-data-exploration

Power Spectrogram

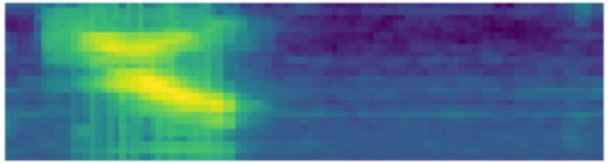
go



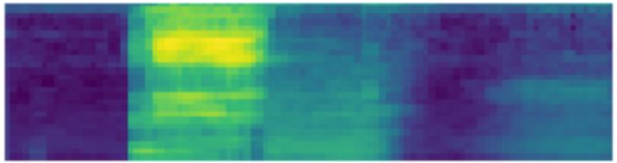
no



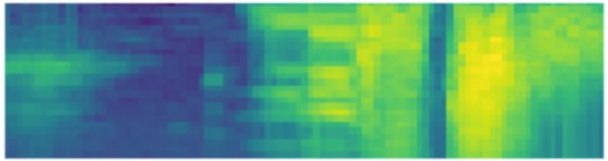
right



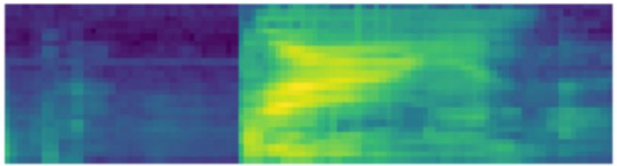
off



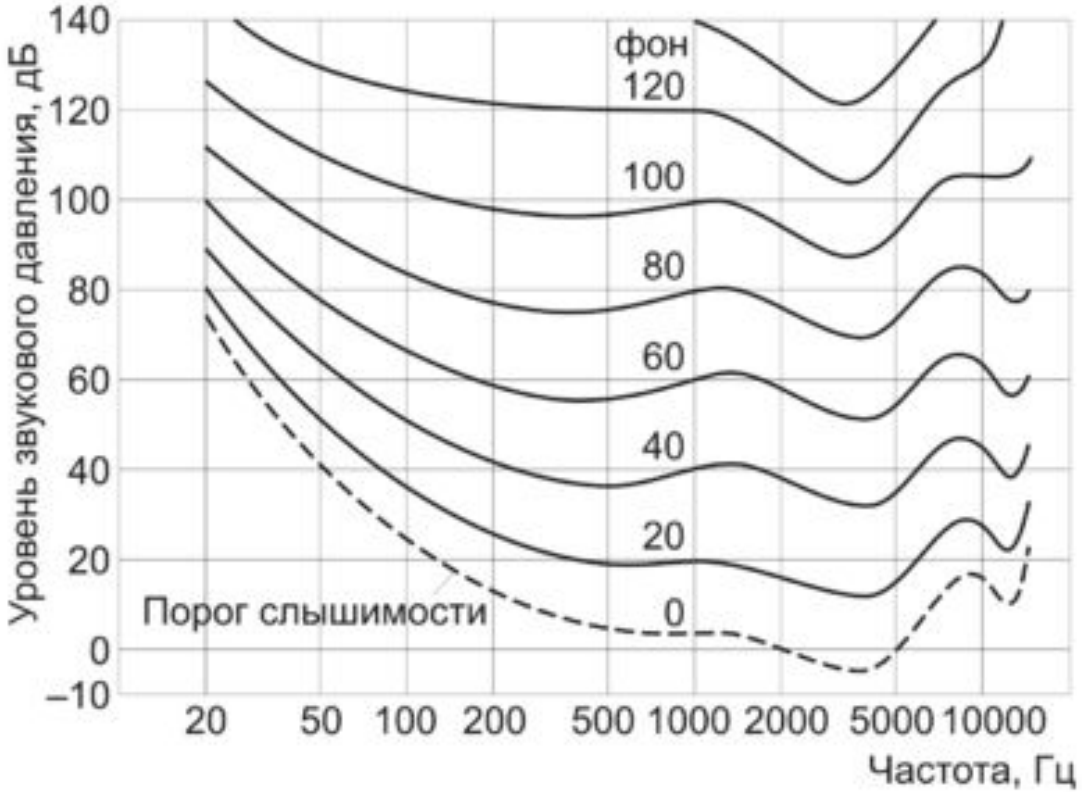
left



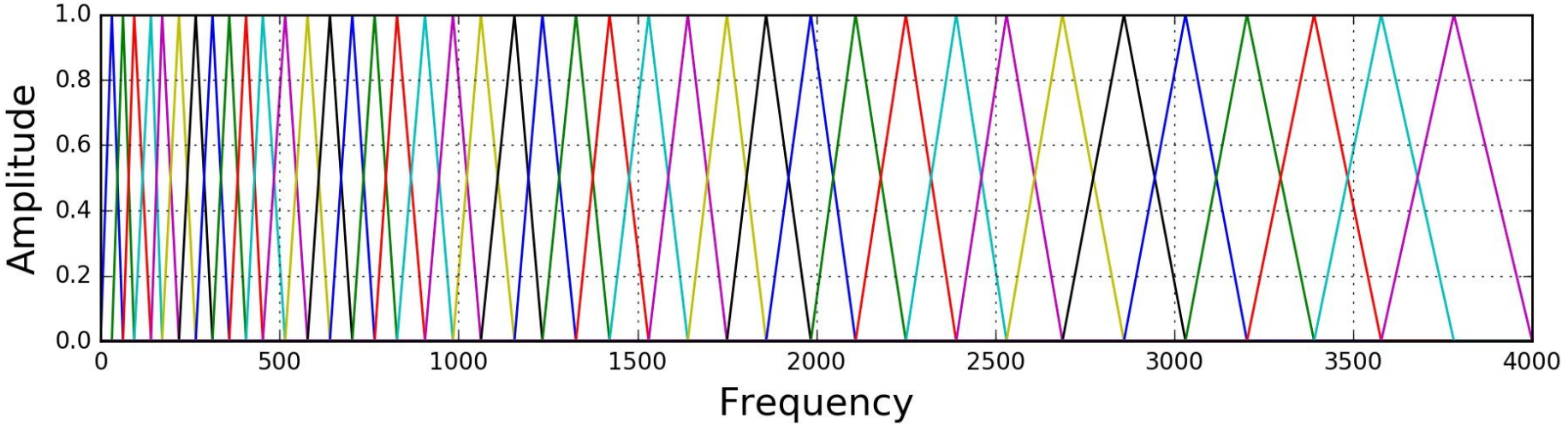
down



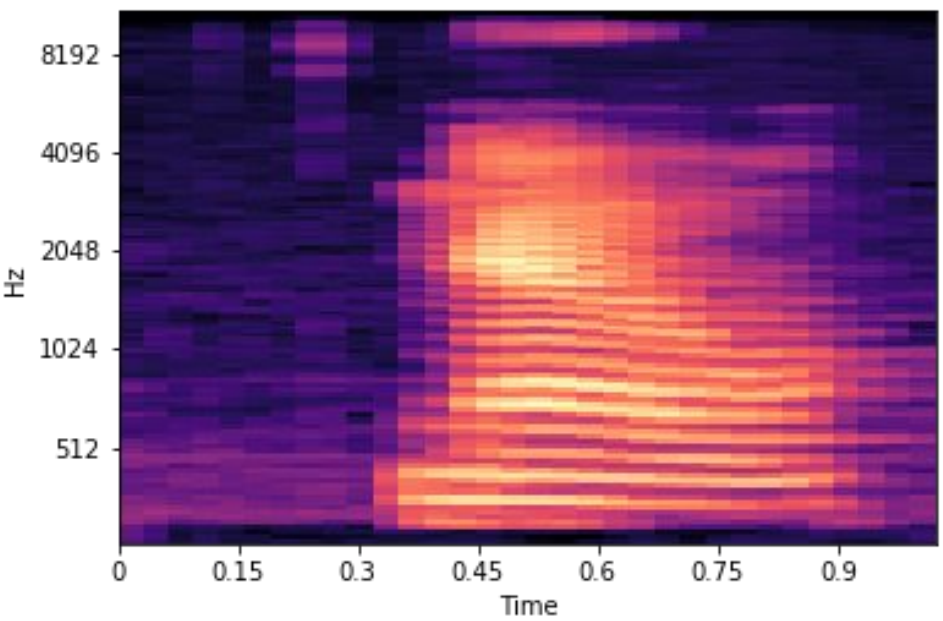
Мел

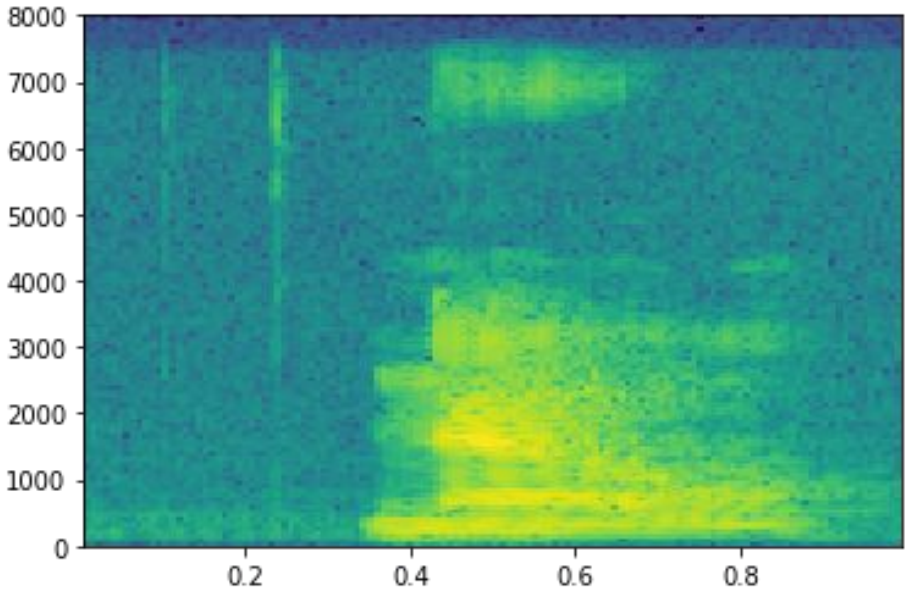

Уровень звукового давления, дБ
140
120
100
80
60
40
20
0
–10
фон
120
100
80
60
40
20
0
Порог слышимости
20
50
100
200
500
1000
2000
5000
10000
Частота, Гц

Mel spectrogram

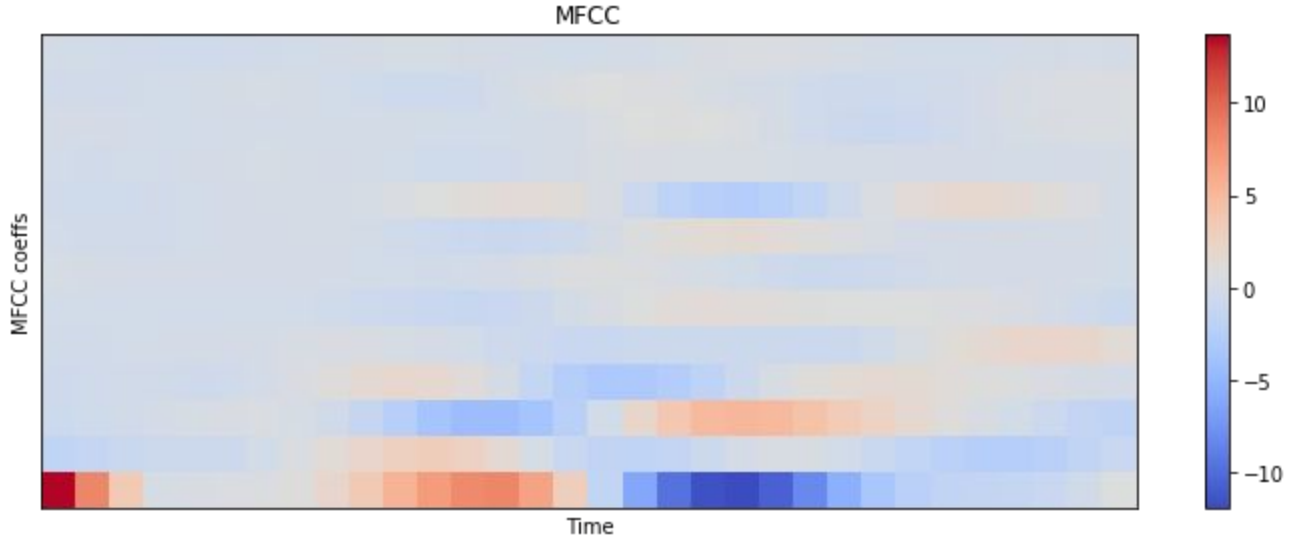


Mel spectrogram

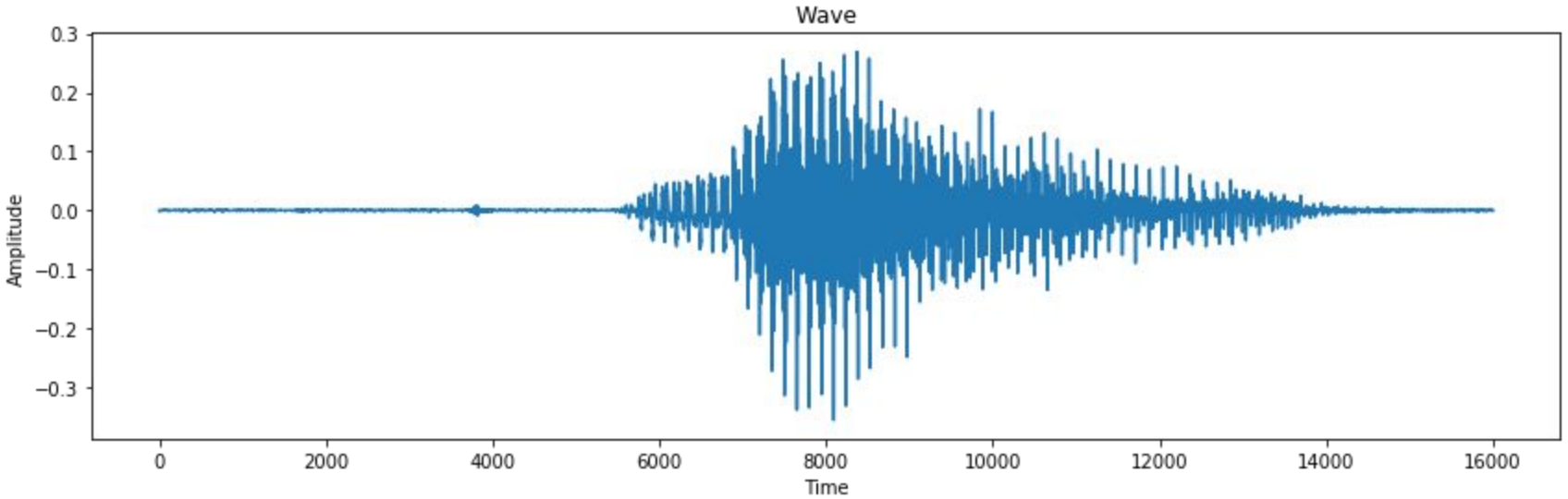




Mel-frequency Cepstral Coefficients



Чистый звук


Wave
Amplitude
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
2000
4000
6000
8000
10000
12000
14000
16000
Time

Чистый звук

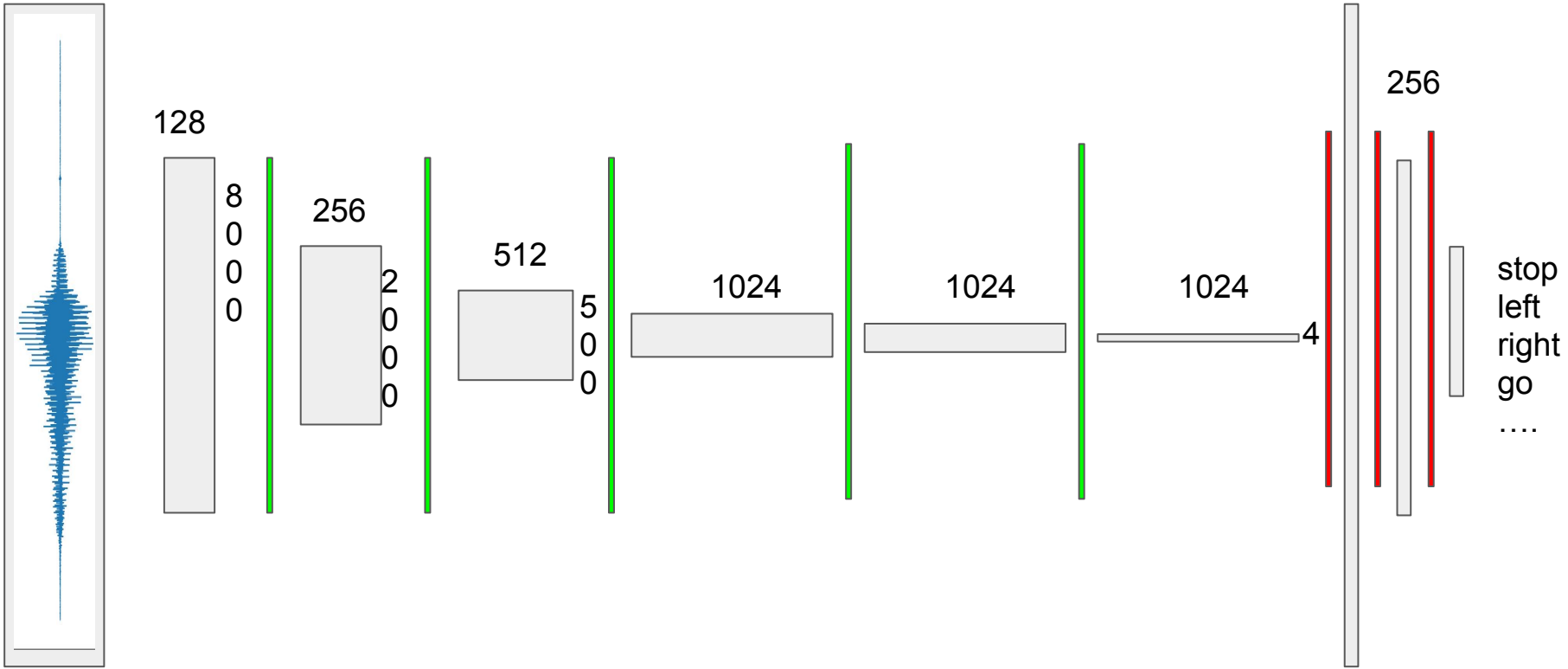
Подготовка звука

- Pre-emphasis: $y[n] = x[n] - \alpha x[n-1]$
- Нормализация громкости:
  - MinMax -> [0..1]
  - Из-коробки librossa для np.float [-1..1]
  - …
- Удлинение коротких клипов (были баги в данных)

Аугментация

- Добавление шума (белый, розовый, автострада, кухня)
- Случайные сдвиги во времени (~200мс)
- Растягивание во времени (!)
- Случайные растягивание по амплитуде (?)

Первое вхождение в сон ~ 0.75 accuracy



Добавим ещё слоёв?

(Нет)

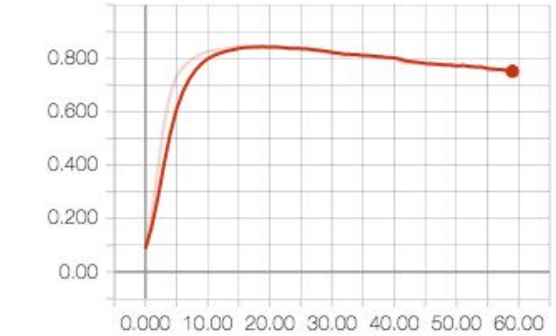
Оно не обучается

Какой оптимизатор использовать: Adam, SGD + nesterov, RMSProp, Adadelta… тысячи их

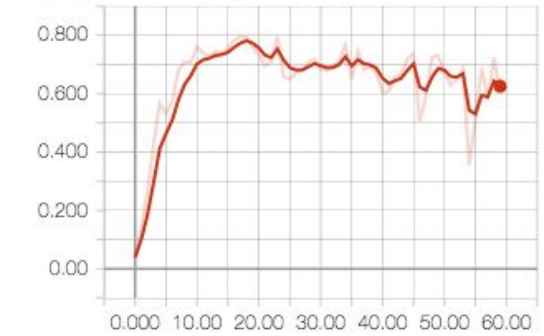
Как подобрать learning rate для оптимизатора

Как подобрать learning rate decay

categorical_accuracy

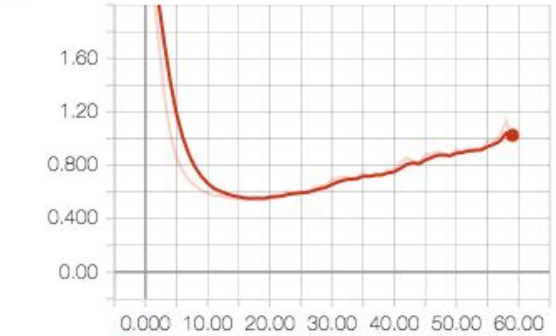


val_categorical_accuracy



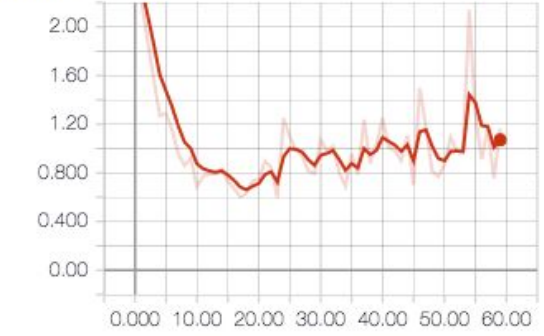
loss

loss



val_loss

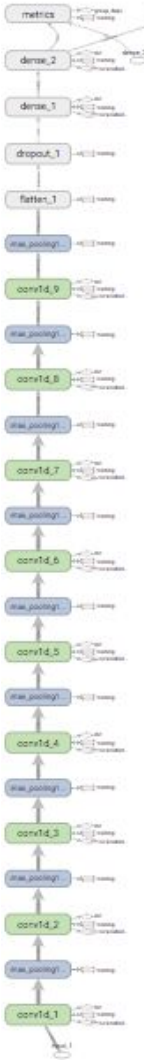
val_loss



Новые слова

Batch normalization

Kernel_regularizer

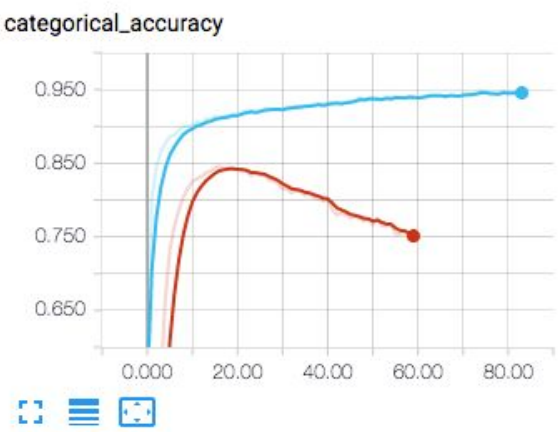


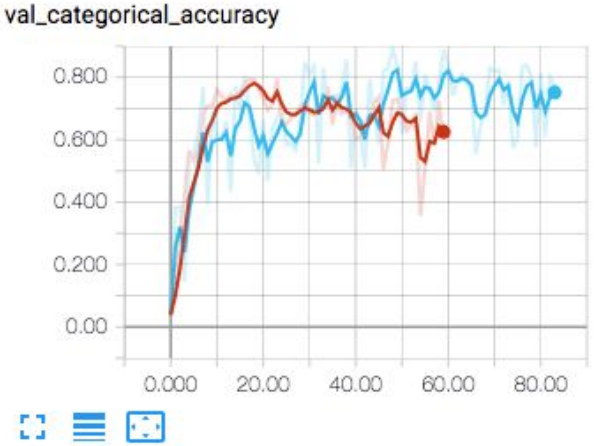
Было:

- 6 свёрточных (Relu)
- Flatten
- 3 полносвязных

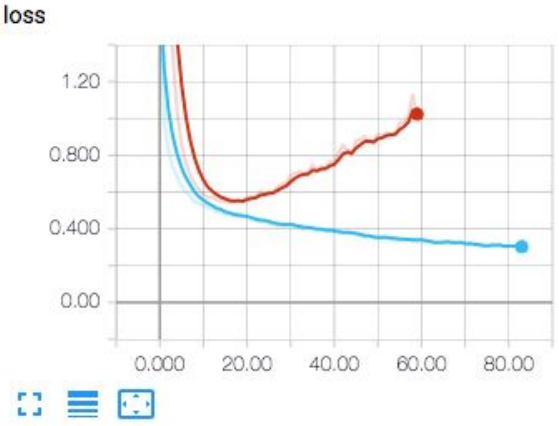
Стало:

- 17 свёрточных + BatchNorm + L2 (Relu)
- Global Pooling
- 1 полносвязный (softmax)

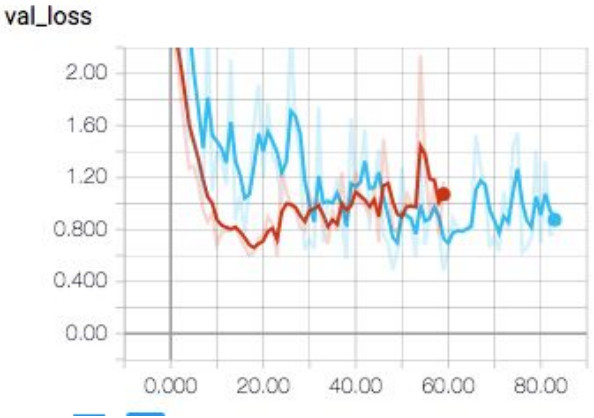




loss



val_loss



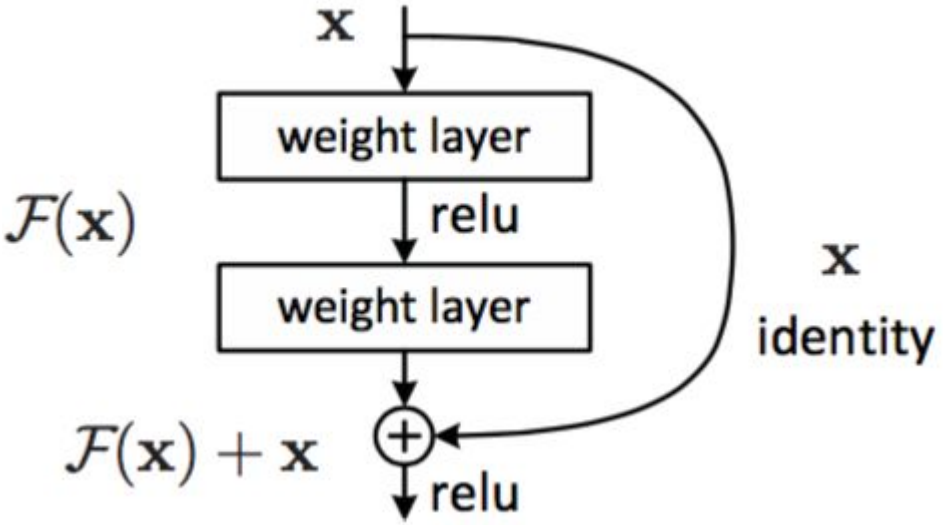
Public 0.8


WE NEED TO GO
DEEPER

Снова новые слова

ReduceLROnPlateau

Residual Networks



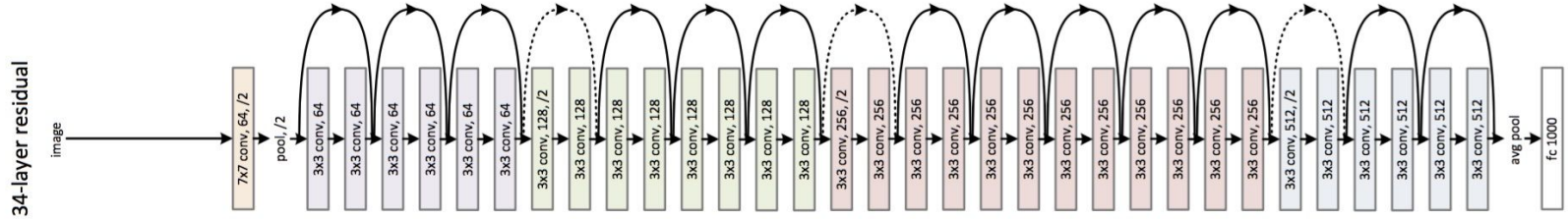
1D ResNet

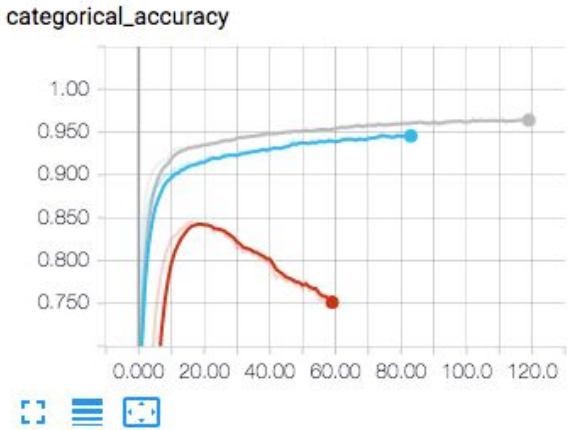
17 Residual блоки по 2 свёртки внутри

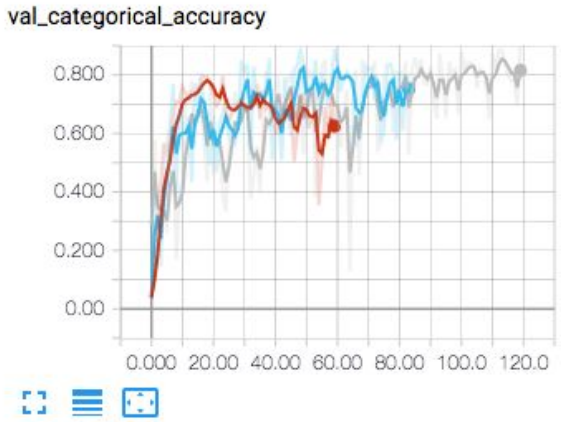
Итого 34 свертки

GlobalAveragePooling

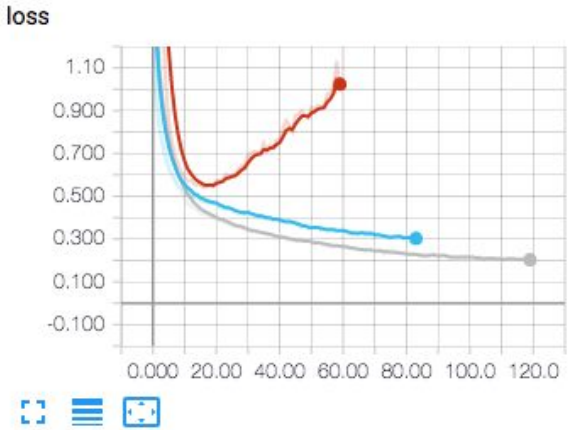
1 Dense слой на выход (softmax)



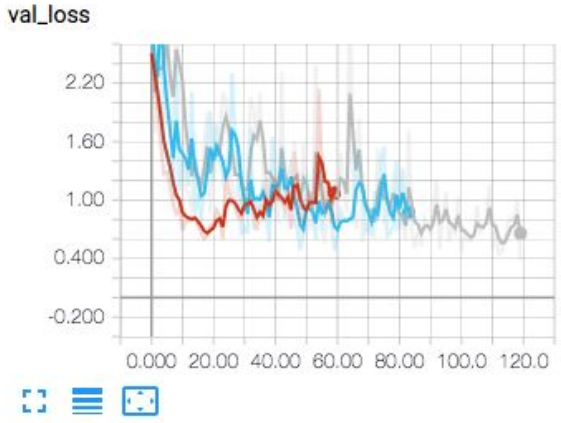




loss



val_loss



Public 0.82

Голосование моделей

Три лучшие модели на validation ~ [0.79..0.82]

Обычный HardVoting

Результата на private: 0.84

Confusion matrix

	down	go	left	no	off	on	right	stop	unknown	up	yes
down	245	3	0	11	0	0	0	1	4	0	0
go	10	204	1	14	0	1	2	7	18	1	0
left	0	0	230	6	0	0	0	1	7	0	3
no	1	3	4	250	0	0	0	0	11	1	0
off	0	1	1	0	222	3	0	1	4	23	0
on	1	0	0	1	9	221	2	4	17	2	0
right	0	0	8	0	0	0	226	0	20	1	0
stop	0	3	3	0	3	0	0	229	6	2	0
unknown	52	30	67	65	20	24	51	30	3838	20	8
up	0	1	0	0	4	1	0	4	7	242	0
yes	1	0	17	5	0	0	0	3	2	0	233

Литература

1. https://github.com/blan4/kaggle_speech_recognition - исходники
2. https://www.kaggle.com/c/tensorflow-speech-recognition-challenge
3. https://www.youtube.com/watch?v=UMh9EmgkN6w
4. http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf
5. http://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html
6. https://habrahabr.ru/post/140828/ MFCC
7. https://arxiv.org/abs/1512.03385 ResNet
8. https://arxiv.org/abs/1710.06554 Honk - решение организаторов
9. http://deeplearning.net/wp-content/uploads/2013/03/pseudo_label_final.pdf PseudoLabeling