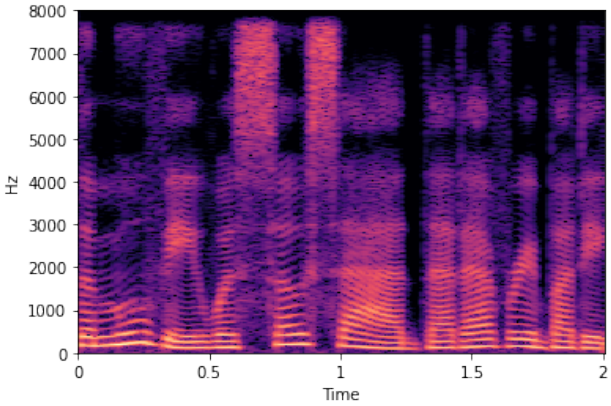
Lime Experiments
Inference Significant Classification Criteria

Wav spectogram : (Real) file2.wav



Conclusion:

- The Lime explanations of masks follow a normal distribution for the green and blue area, this concluded ploting many histograms, two of them are in figures a, b.
- Considering this proximity of distribution of pixels we can give explanation of the color areas more safe.using ht erule of thumb then majiority of green color is in 2sigma on y axis which correspond in probability of 97% to meet this pattern there.

The area for explainability from the experiments 2 to 4 and experiment 5 on figures 7, 8 ranges from medium robustness to more robustness from left to right as you read the page.
From the experiments, on frequency (y axis) we see the band of frequencies between (6960, 1040) Hz in voice one of the most important for patterns of the Neural Network.
On time axis the pauses in tonality shows the largest contribution because many the green areas exists there.
On y axis the smoothness/ sharpness of the voice contributes too because on fake dataset the voice is more stip between pauses resulting in less expand of green regions horizontaly.

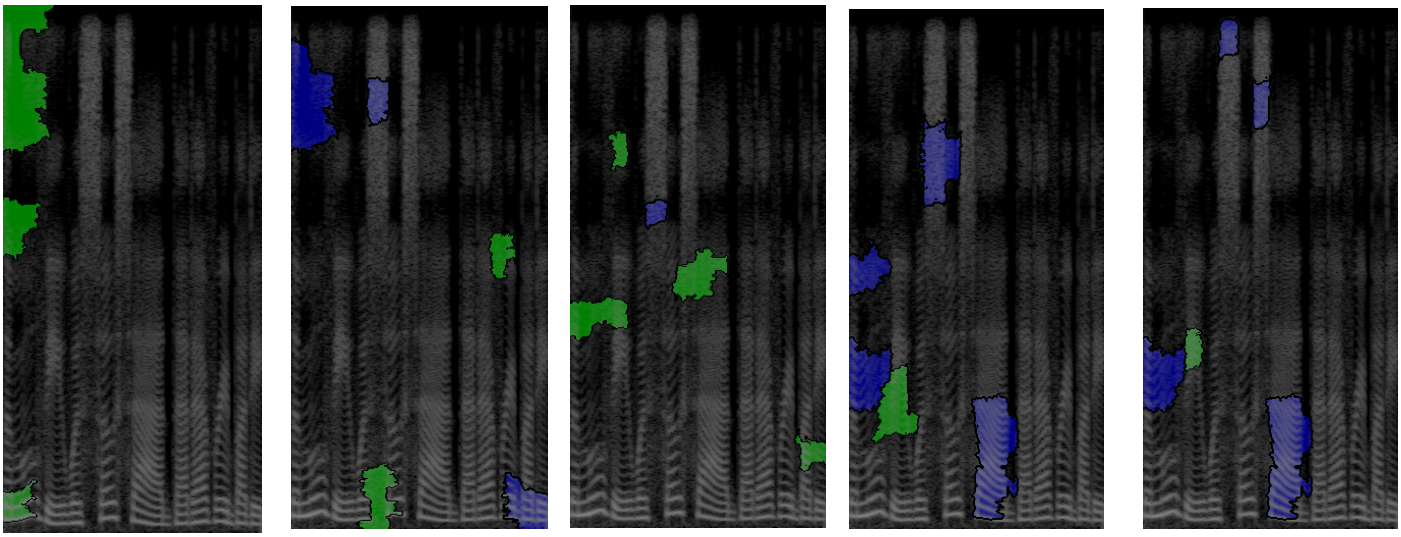
- Between real and fake recognition we can observe that the robustness drops for the fake given the same values on hyperparameters, and the model looks a wider range of frequencies to inference a result. A legit explanation is that the fake voice is more uniform making it more difficult to extract patterns from certain regions.

Description of important hyperparameters

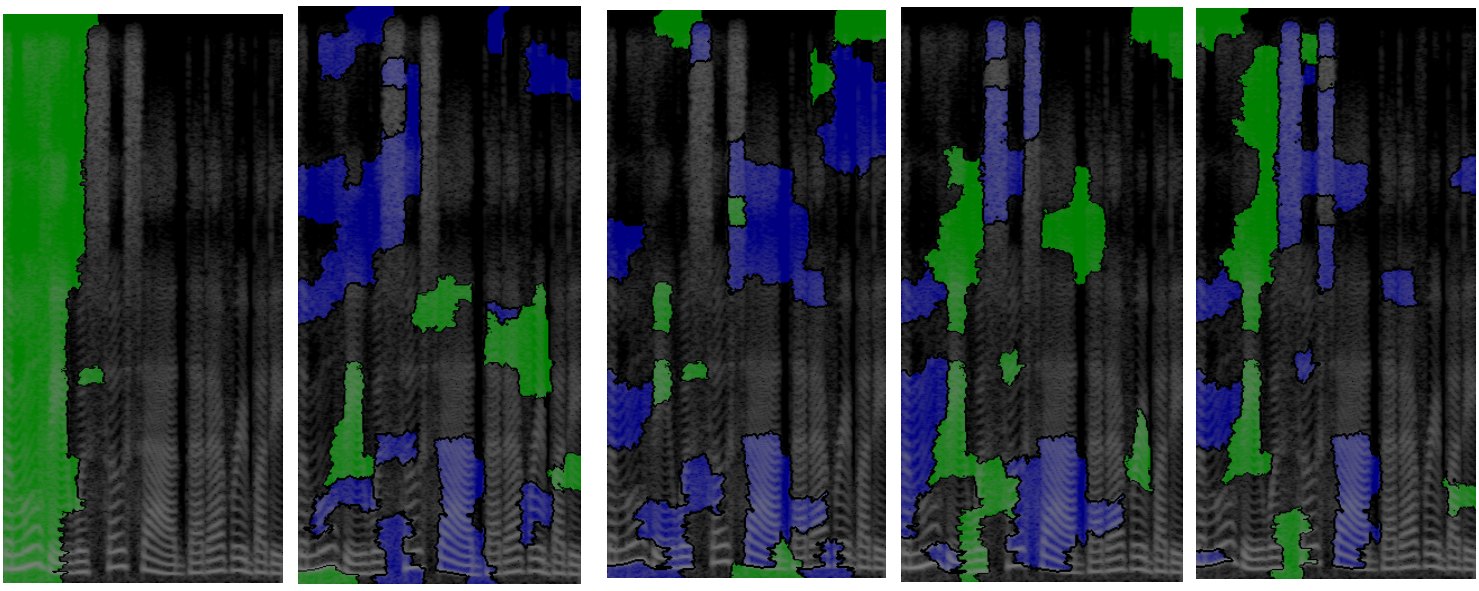
- kernel_width:
  The affect of smoothing in a pixel's neighboor
- top_labels:
  The number of classification labels
- hide_color:
  The type of colors
- num_samples:
  With larger n_samples it takes more CPU time and RAM to explain a prediction, but it could give better results. Larger n_samples could be also required to get good results
- Number of features:
  The lower, the easier it is to interpret the model. A higher potentially produces models with higher fidelity aka is how reliable the interpretable model is in explaining the black box predictions in the neighborhood of the data instance of interest.

- Metric: Weighted R^2
  Robustness of the model affected by kernel_width. Values range (0, 1)

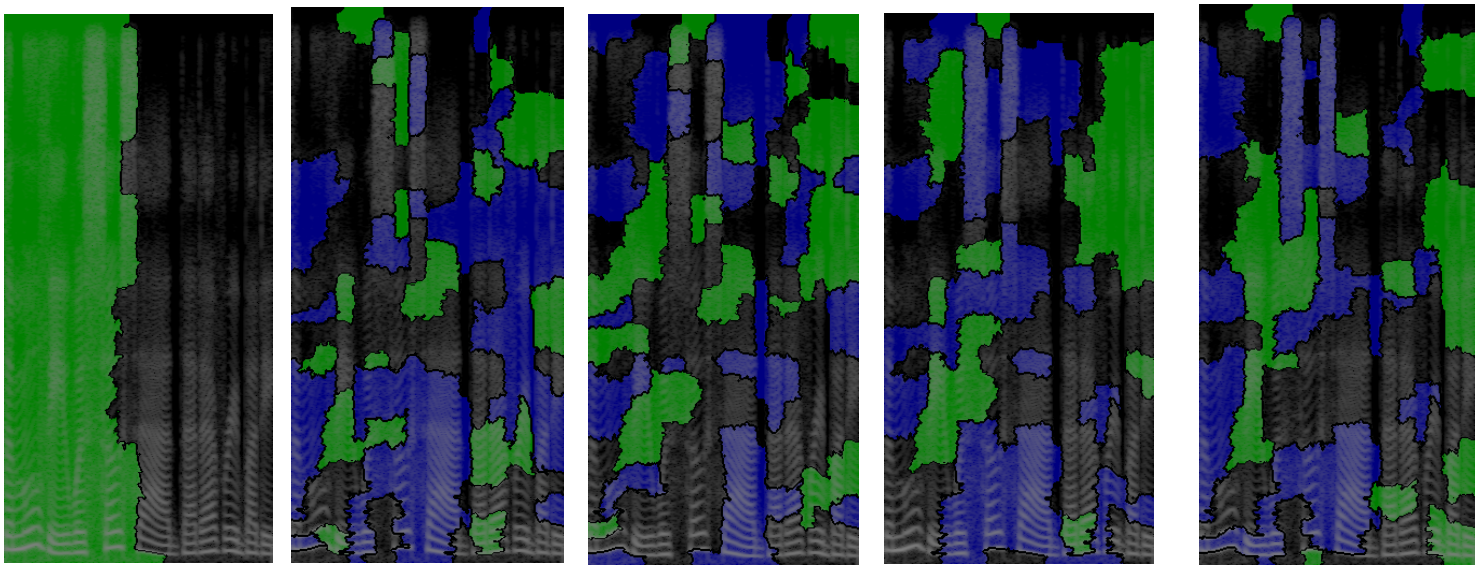
Experiement 1
top_labels=2 | hide_color=2 | num_samples=5, 15, 50, 500, 1500 | num_features=5



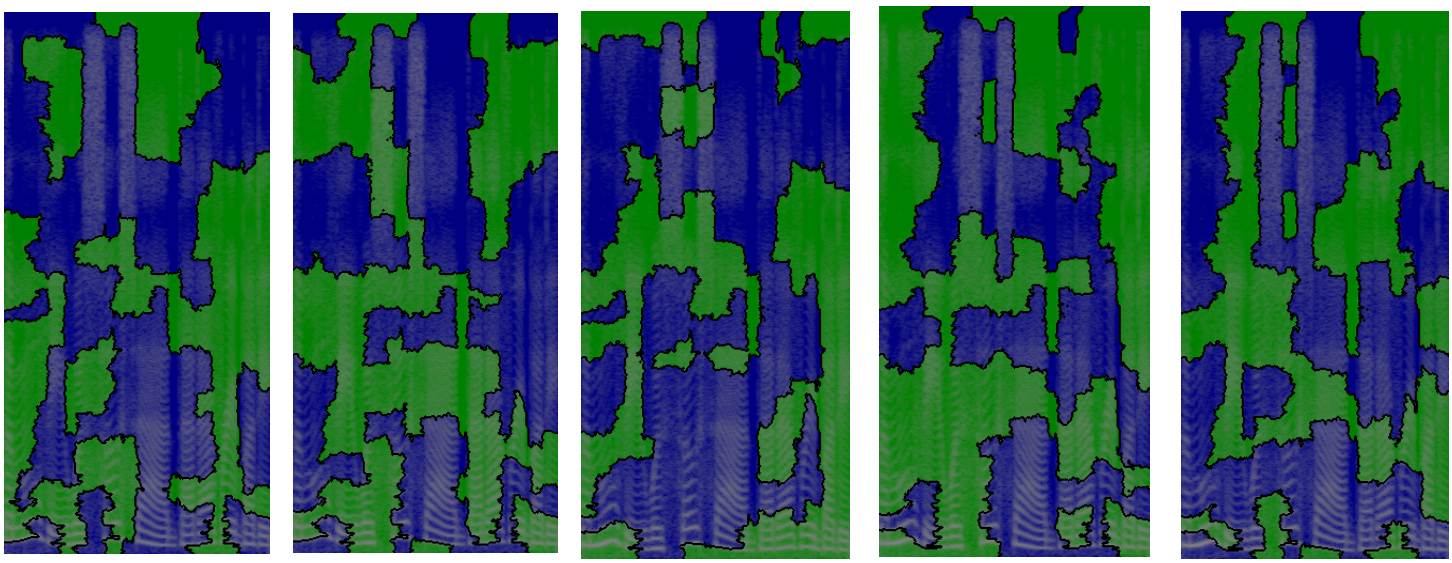
Experiement 2
top_labels=2 | hide_color=2 | num_samples=5, 15, 50, 500, 1500 | num_features=25



Experiement 3
top_labels=2 | hide_color=2 | num_samples=5, 15, 50, 500, 1500 | num_features=50



Experiement 4
top_labels=2 | hide_color=2 | num_samples=5, 15, 50, 500, 1500 | num_features=150



Experiement 5

top_labels=2 | hide_color=2
kernel_width=0.07, **0.07**, 0.2, 0.2, 0.8**,** 3.0 | num_samples=1.000, **10.000**, 50, 10.000, 10.000, 5.000 num_features=15

Shown R2=0.167, **0.117**, 0.96, 0.192, 0.181, 0.202

For large kernel=0.8, 3 (figures 5 and 6) despite the large number of samples (10.000, 5.000) the model looses roboostness (R2=0.181, 0.202), also for smaller number of samples (50 samples) looses roboostness.

- Between kernel_width 0.07-0.1 is a good range and for large number of samples 5.000-10.000 but the feature are not enough (15) so 50 features is a good choice.

- In Figure 7 increased the num of features to 50, the rest kept the same at the best values | kernel_width=0.07 | samples = 10.000. Shown R2=0.151

- In figure 8 increased the num of features to 150, the rest kept the same at the best values | kernel_width=0.07 | samples = 10.000. Shown R2=0.094

figure 1

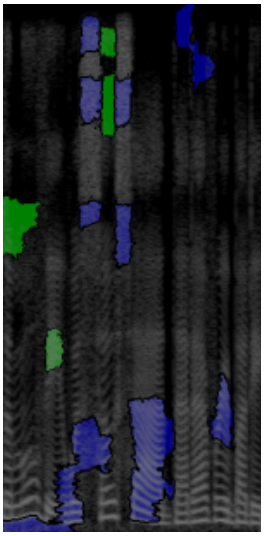


figure 2

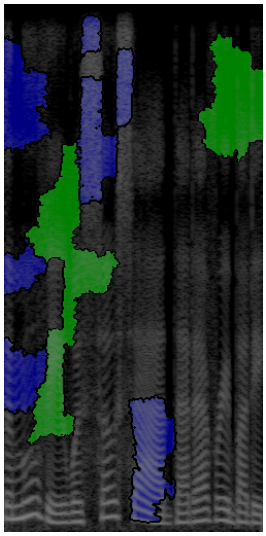


figure 3

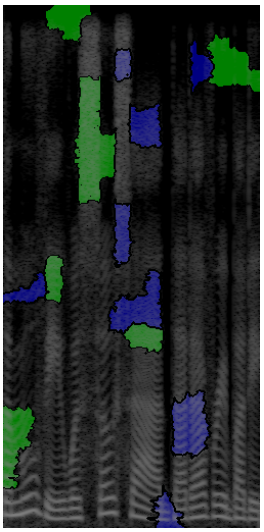


figure 4

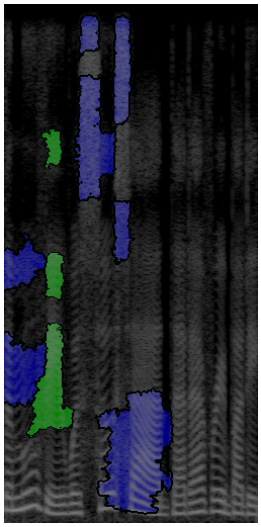


figure 5

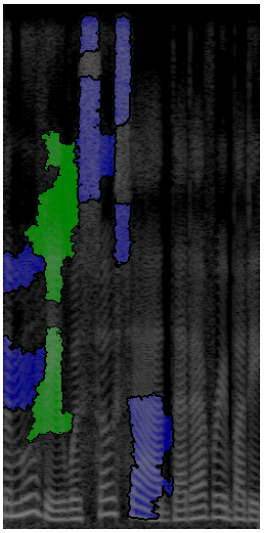


figure 6

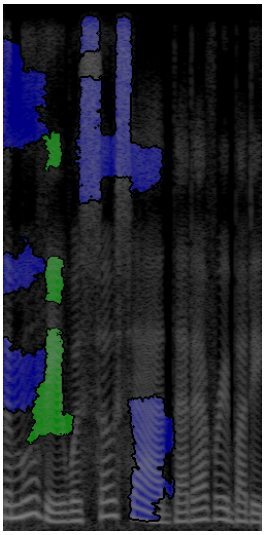


figure 7

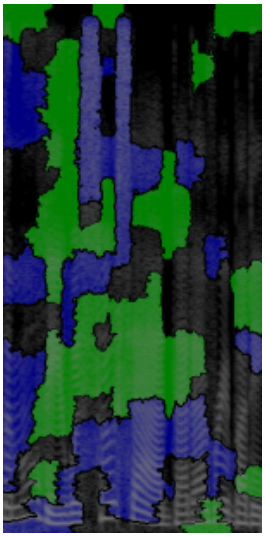
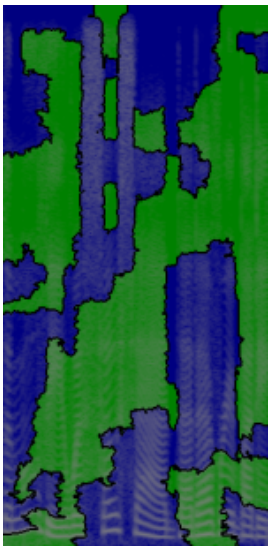
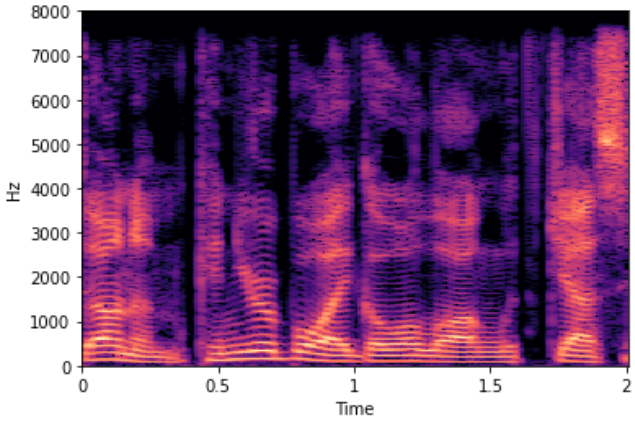


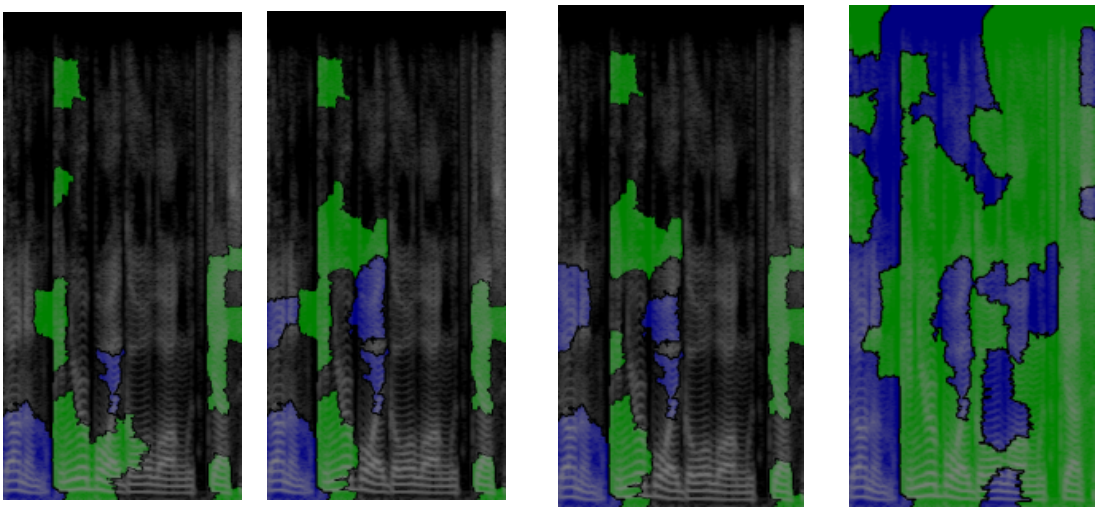
figure 8



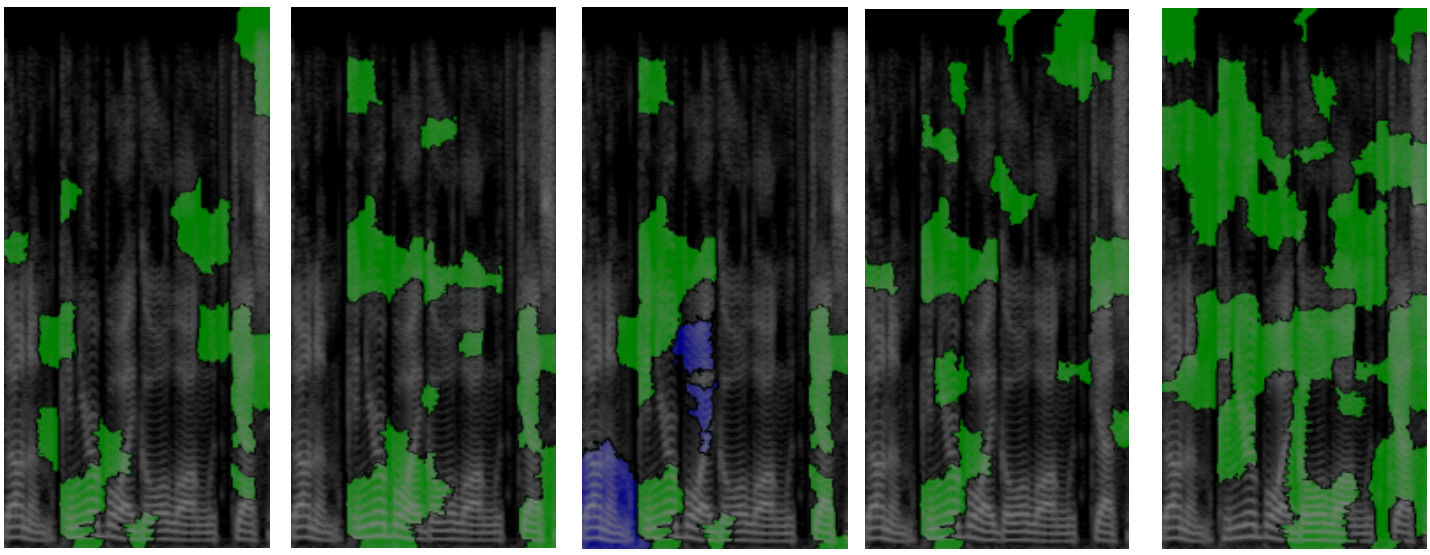
Wav spectogram : (Fake) file6.wav



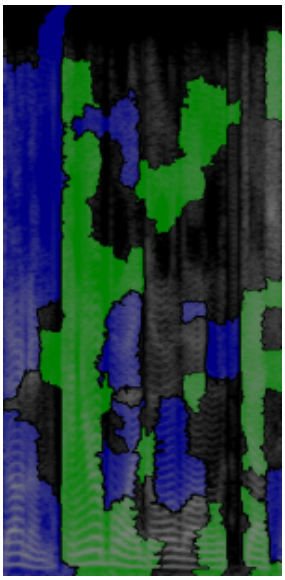
top_labels=2 | hide_color=2
kernel_width=0.07, 0.1 , default | num_samples=10.000, 5000 | num_features= 15, 15, 15, 50
R2=0.440, 0.434, 0.434, 0.503



top_labels=2 | hide_color=2
kernel_width=0.07, 0.07, 0.07, 0.04, 0.04 | num_samples=500, 1.000, 10.000, 10.000, 10.000, 5.000
num_features= 15
R2=0.631, 0.657, 0.541, 0.011, 0.007



top_labels=2, | hide_color=2
kernel_width=default, 0.07| num_samples=1000, 10.000 | num_features=50, 150
R2=0.496, 0.430



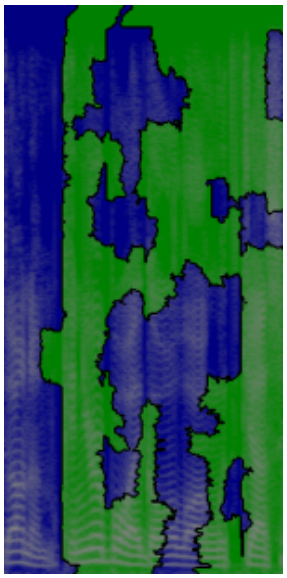


figure a

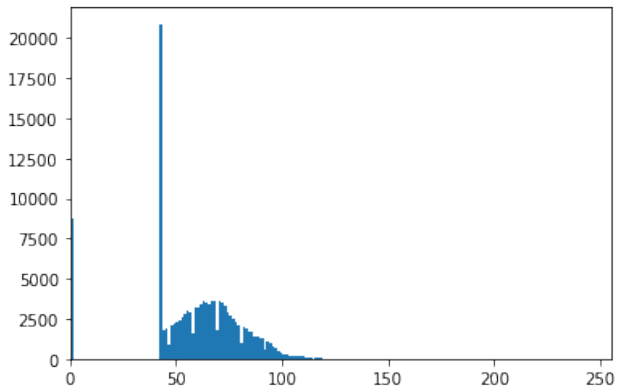
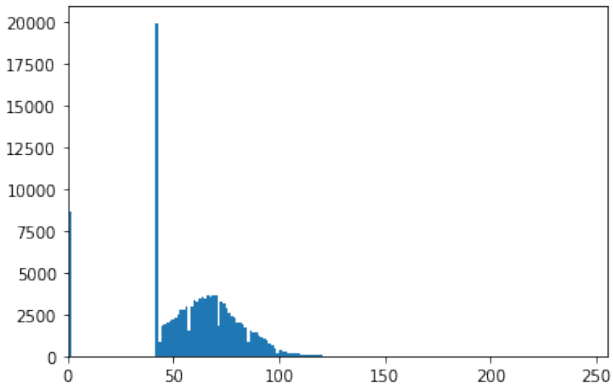


figure b



The x-axis indicates the range of values the variable (color) can take. The y-axis shows the count of how many values fall within that interval or bin.
We can see that the distribution of the pixel intensities is more skewed over the darker sides (0, 40) which is the background color, as the median value is around 70. For the Blue-Green bins the variance is stable because it follows like a Normal Distribution as such we inference the amount of green blue pixels are normaly distributed in the pictures as robustness of LIME model increases. From there the inference of the pictures zones is posible by the human eye with better certainty.

Experiments on 10 real spectograms

wav_real = [
file2.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file9.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file24.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file27.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file30.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file36.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file42.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file43.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file60.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file63.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav]

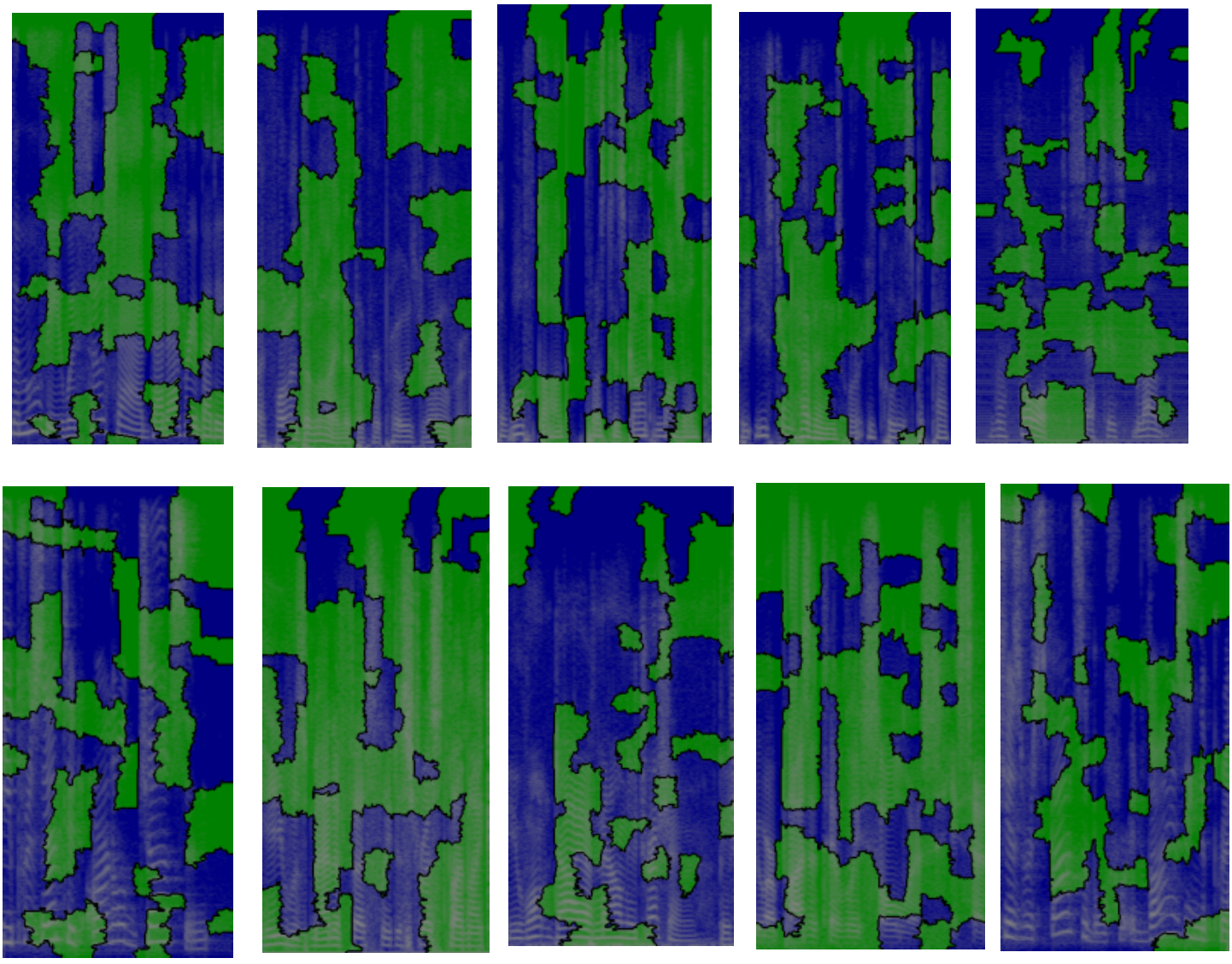
Inference

On real the pauses first then the frequency range (2000-7000 Hz) then the smoothness of the voice play the most significant role for the classification as real.

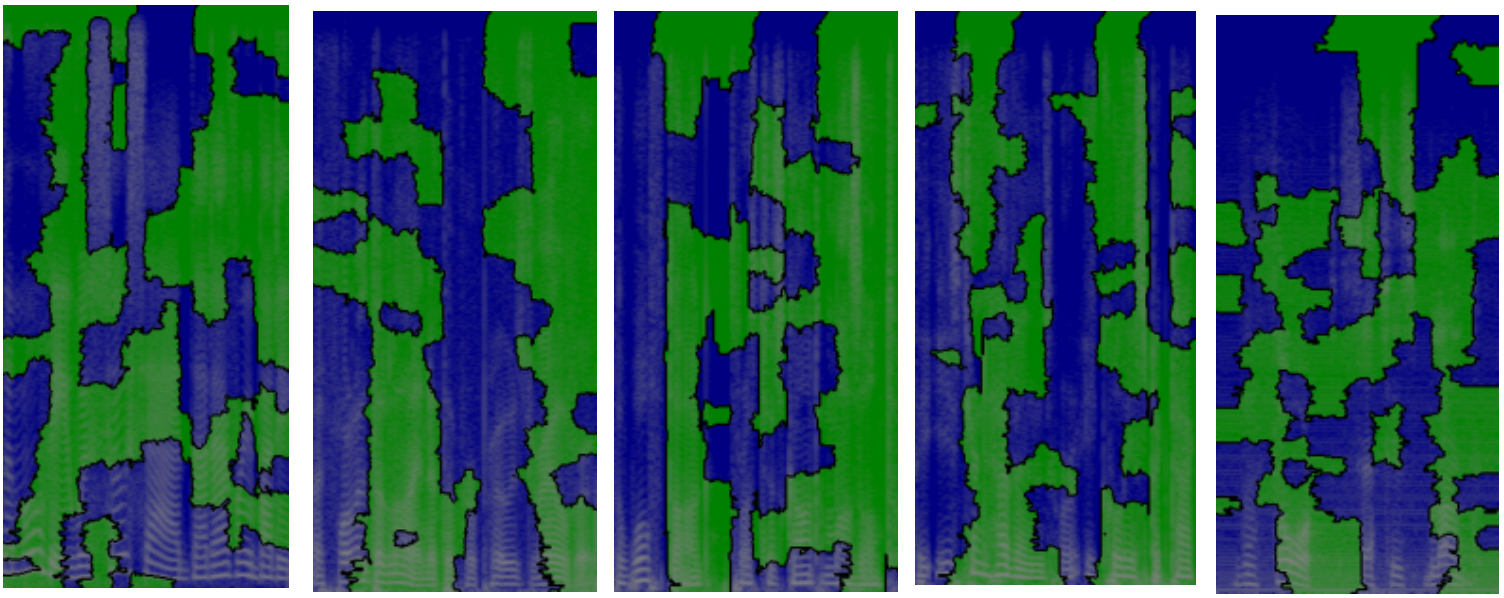
Experiments

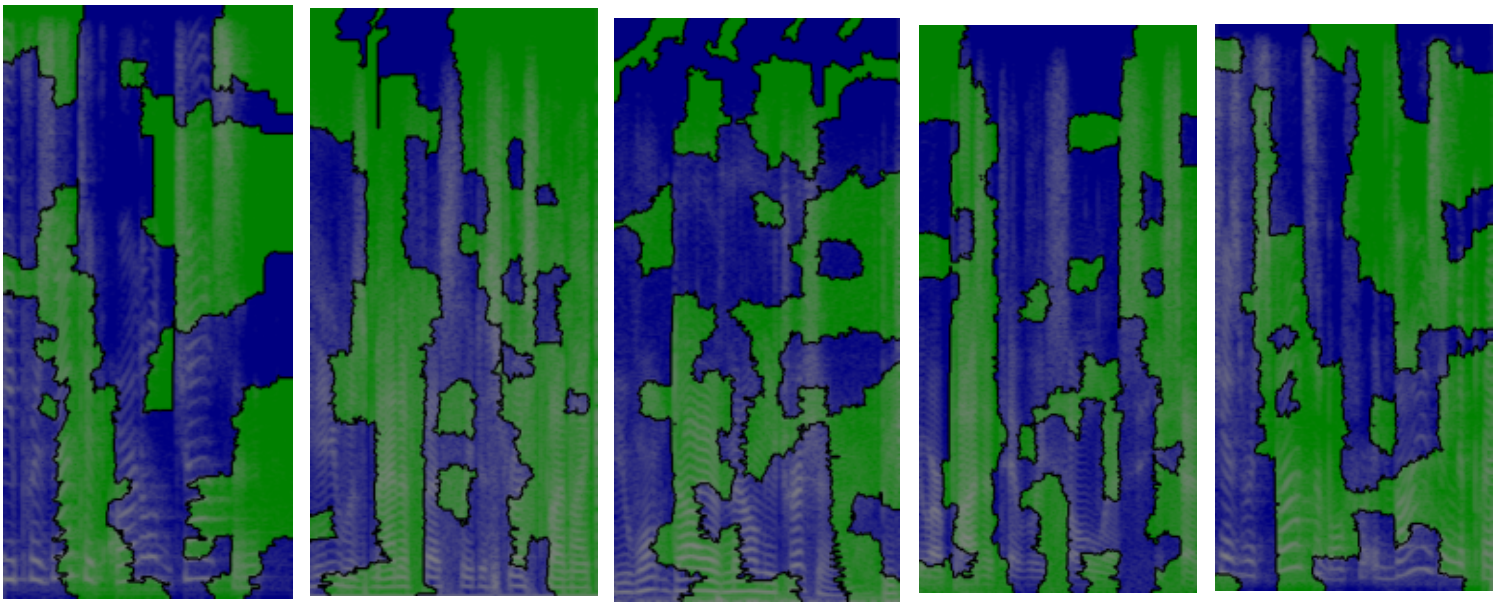
- kernel_width=0.07, num_sampales=1000, num_features=150

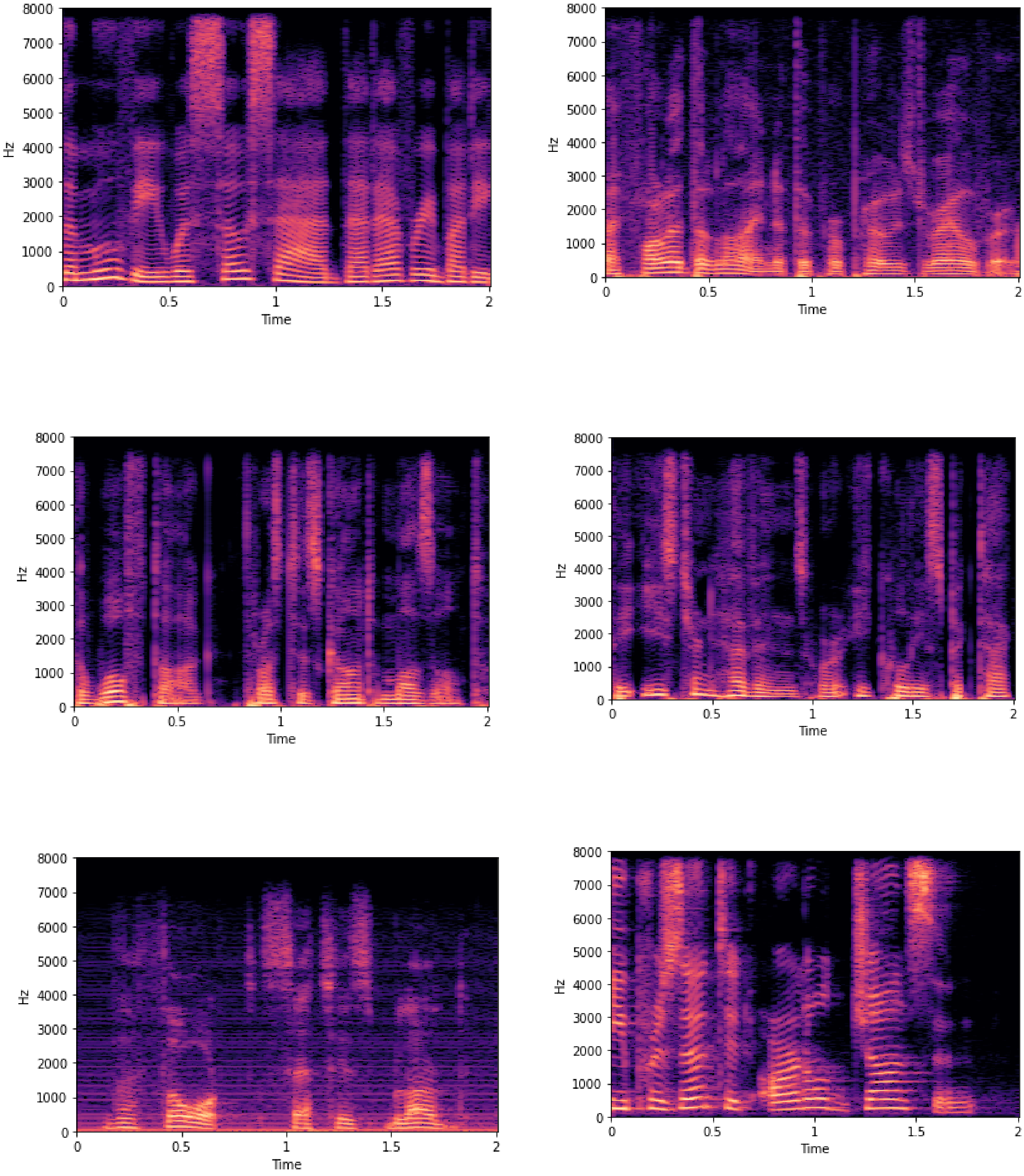
R2 = 0.056, 0.264, 0.344, 0.297, 0.188, 0.158, 0.503, 0.170, 0.339,0.279



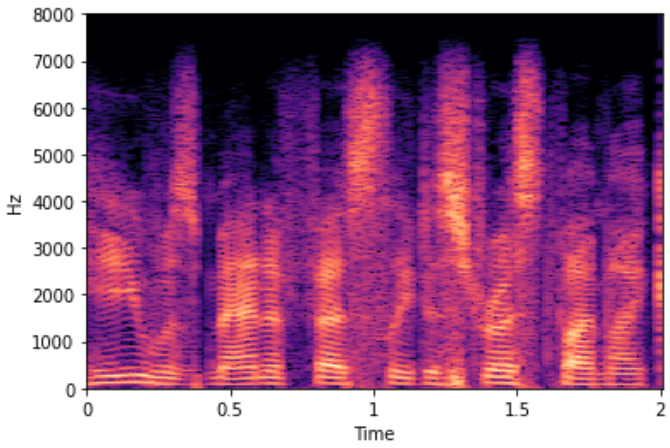
- kernel_width=0.07, sum_samples = 10.000, num_features=150
  R2=0.119, 0.277, 0.266, 0.253, 0.301, 0.225, 0.468, 0.352, 0.43

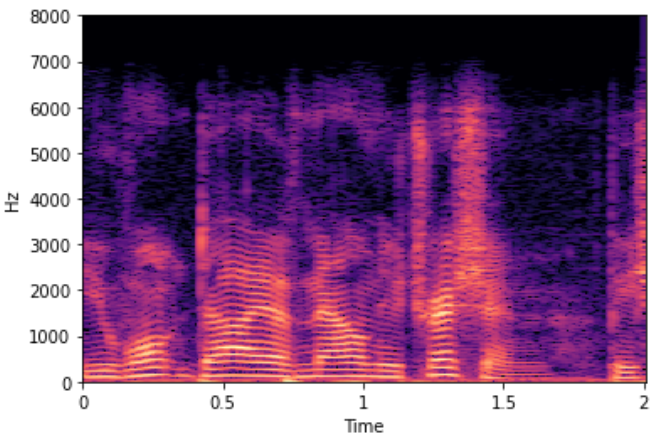


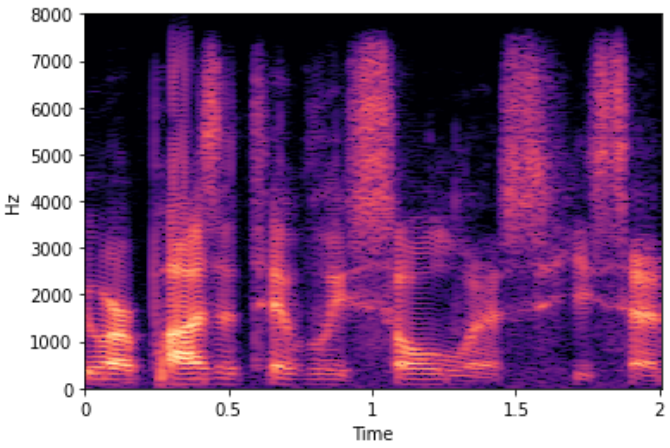


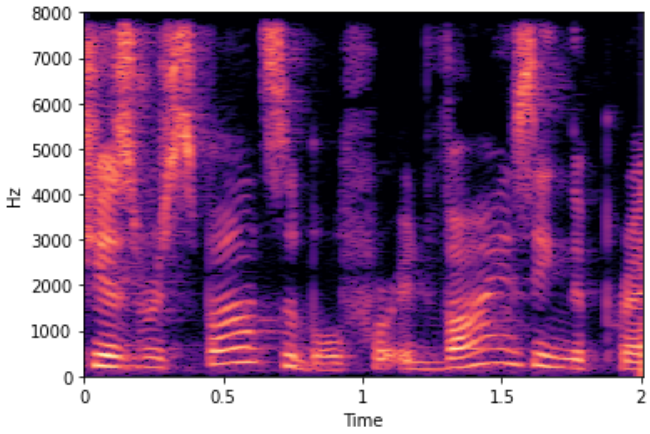


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time

Expereriments on 10 fake spectograms

wav_fake = [
file6.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file7.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file9.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file18.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file26.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file29.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file32.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file33.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file34.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav
file38.mp3.wav_16k.wav_norm.wav_mono.wav_silence.wav_2sec.wav]

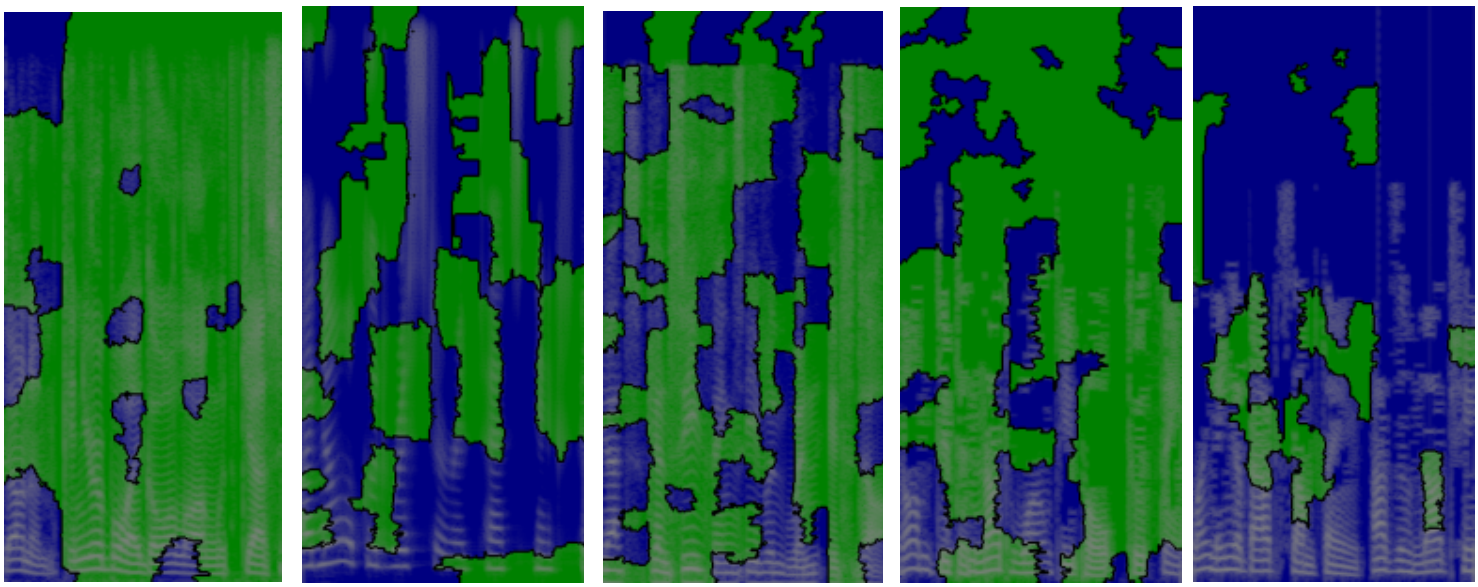
Inference

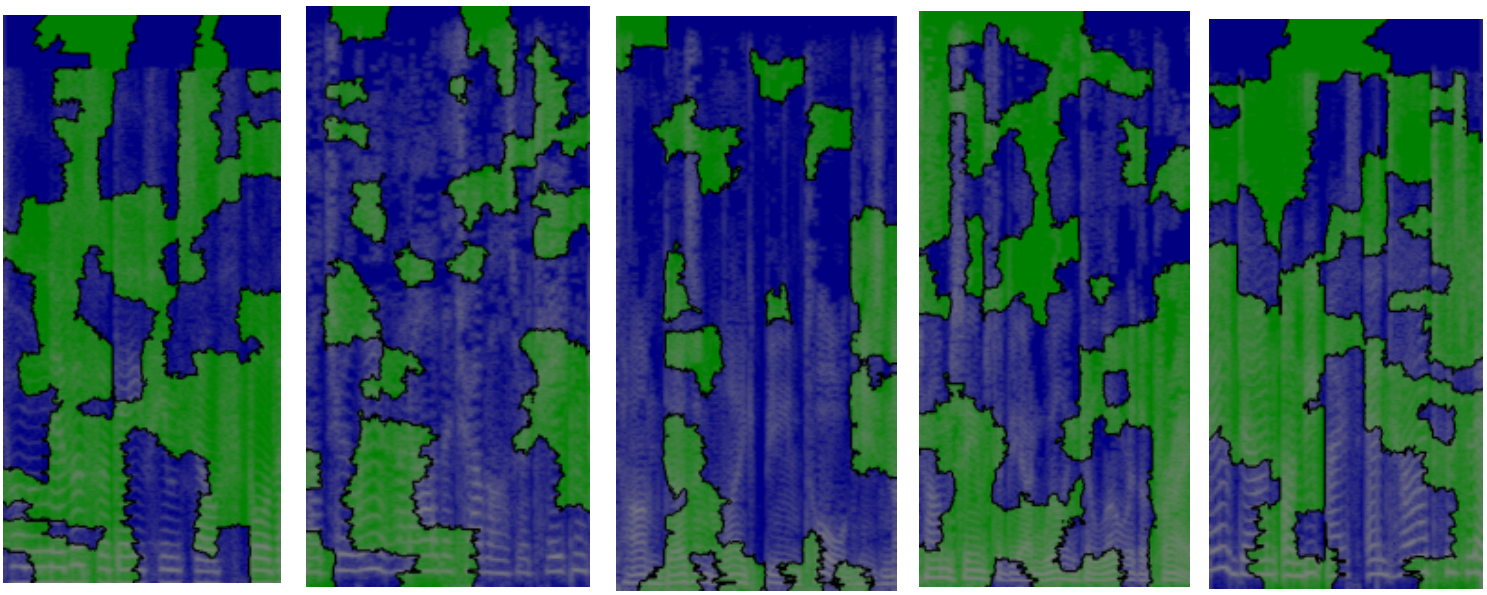
On fake the pauses first then the frequency range (1000-7000 Hz) then the strickness of the voice play the most significant role for the classification as fake.

Experiments

- kernel_width=0.07, num_samples=1000, num_features=150

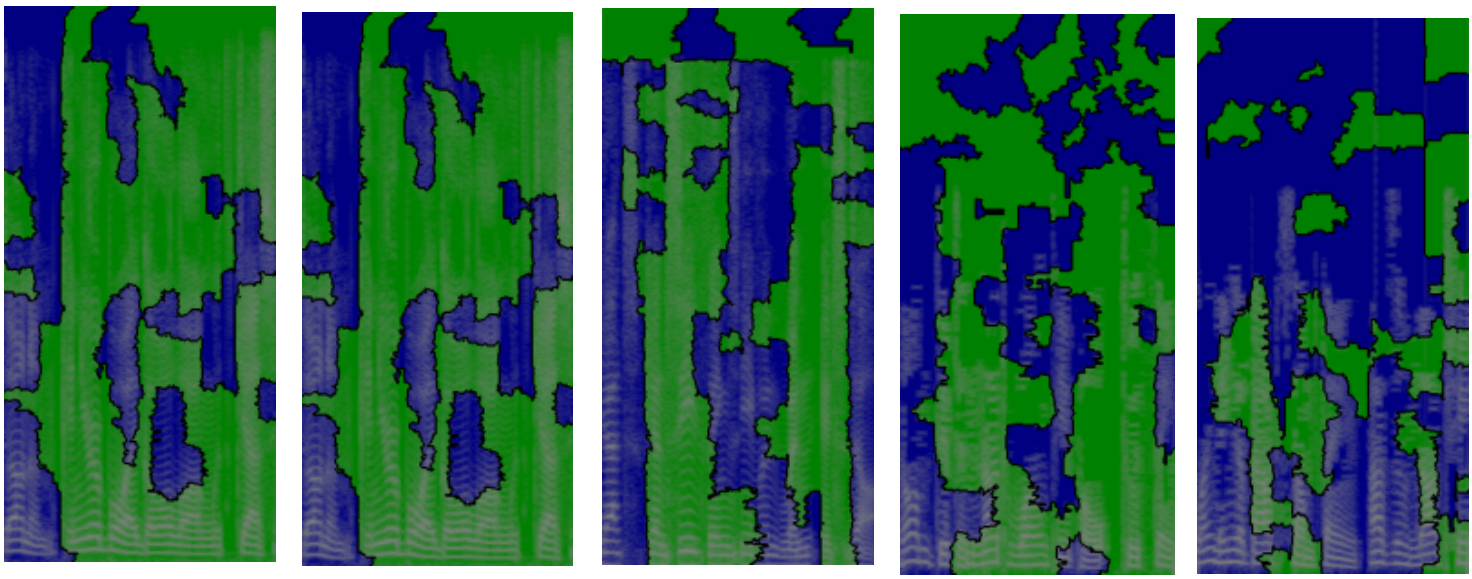
R2=0.660, **0.147**, **0.191**, 0.410, 0.6, 0.227, 0.395, 0.565, **0.044**, **0.027**

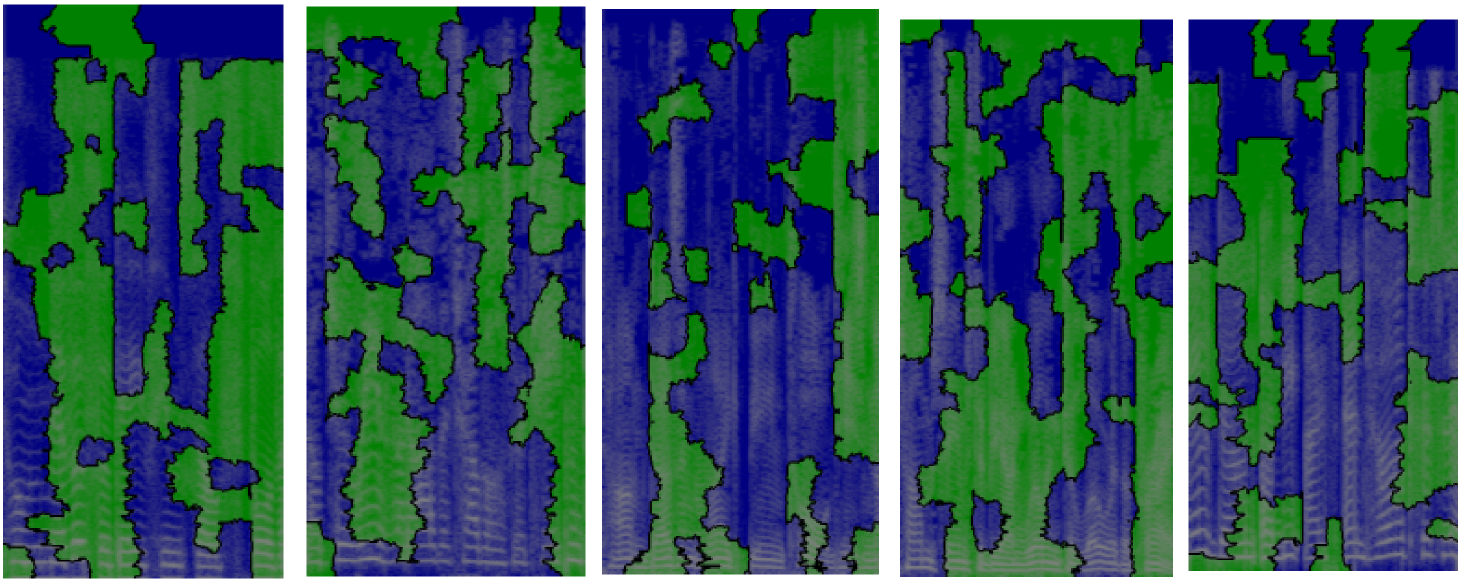




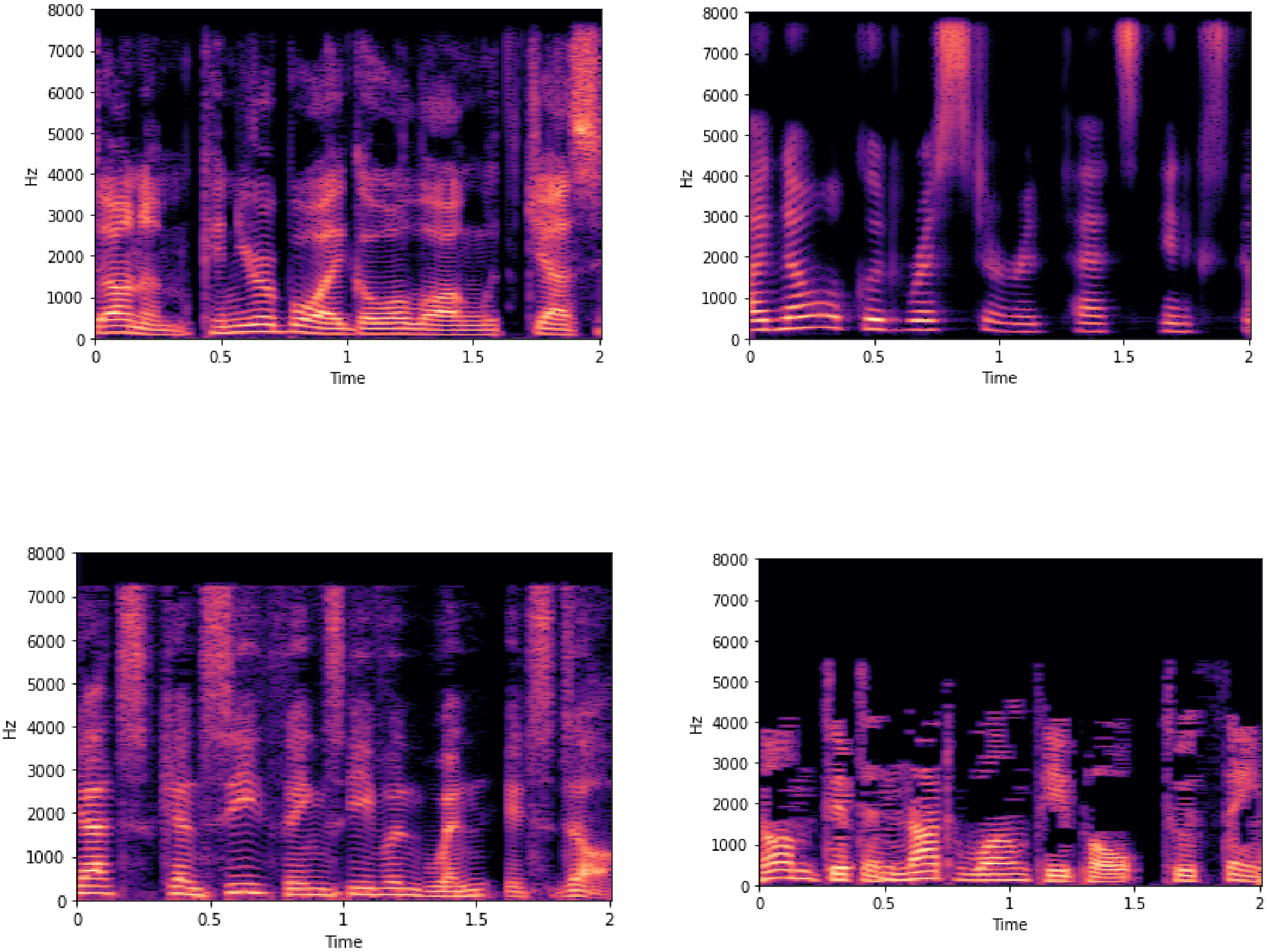
- kernel_width=0.07, num_smples=10.000, num_features=150

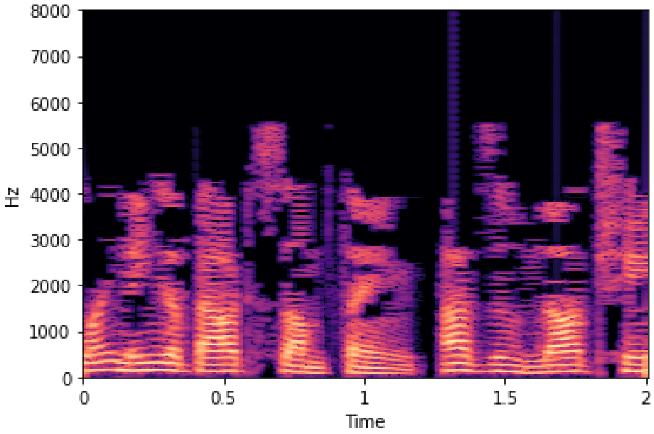
R2 = 0.533, **0.125**, 0.250, 0.464, 0.549, 0.216, 0.429, 0.543, **0.1**, **0.007**

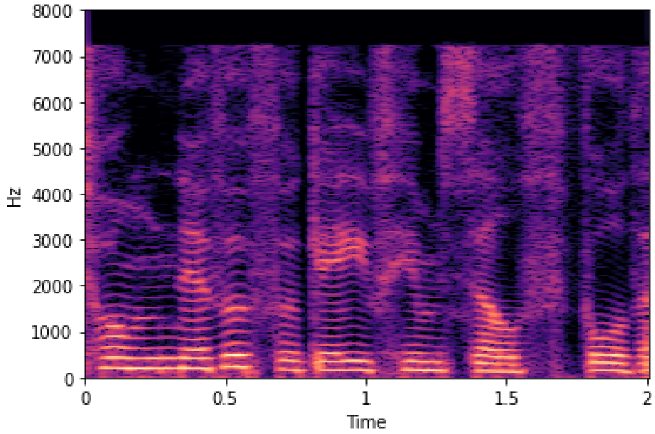


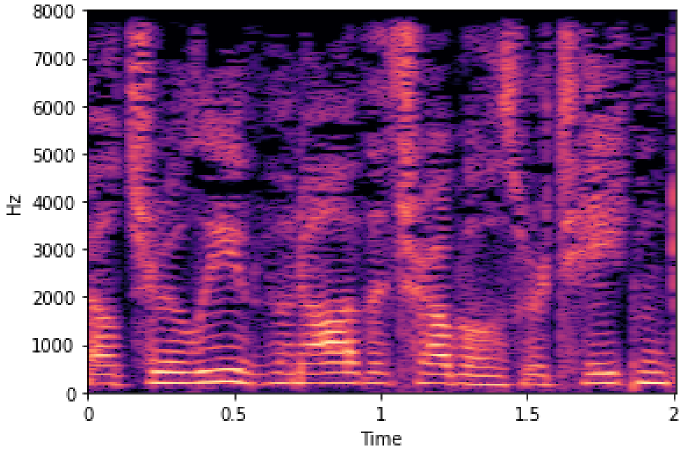


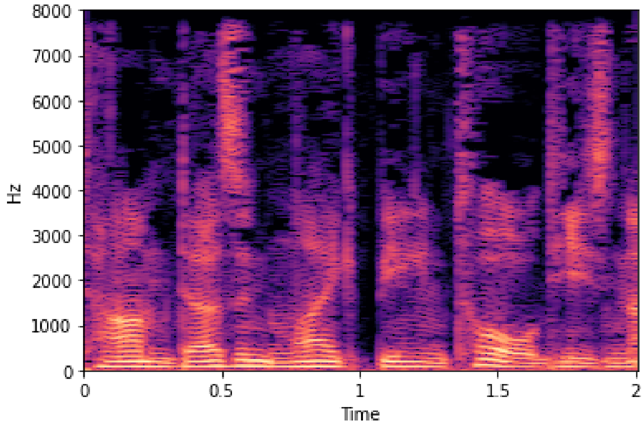
Spectograms (fake)

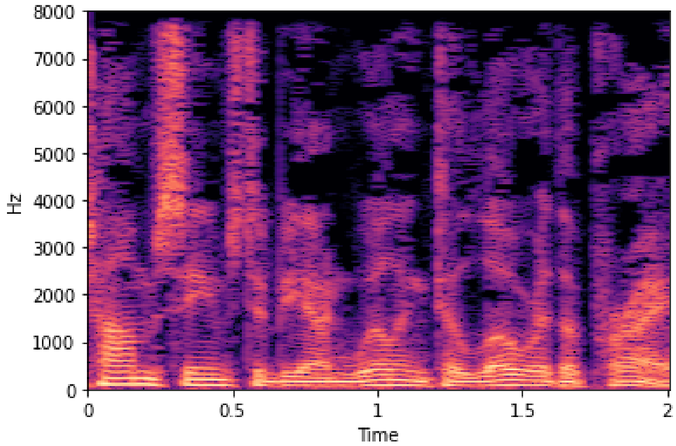


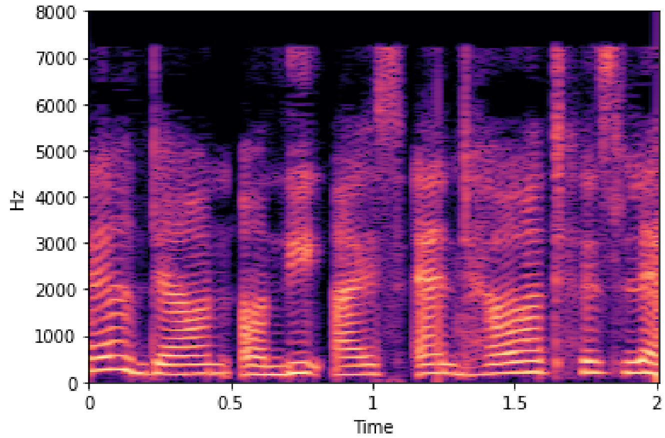

8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time


8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
0.5
1
1.5
2
Time