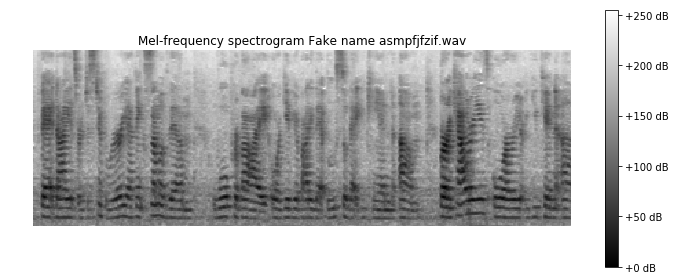
Audio Analysis:

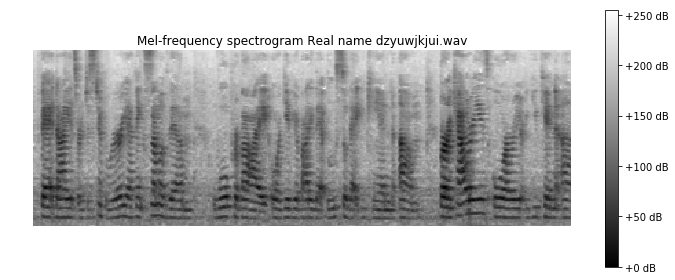
Based on the competition guideline, deepfake videos can be either image or audio modified, while majority of our effort have spent on image classification, it’s worthwhile to exam videos that are manipulated on audios. Unlike image, audios fakes are much harder to discern by human judgement, when listening an example of an audio fake versus it’s original, the sound seems to be the same if not completely identical.

Based on discussion from the Kaggle forum, One way to recognize fake audios is to compare the originals and the fake to identify any mismatch sample rate or missing audio tracks, if any of the audio sample differ, or if the fake video contains audio and the original does not, or if the sample rates do not match, those videos will be considered to have modified audio, conversely, if the video contains no audio or if the sample rate matches exactly, such instance are marked as not containing modified audio.

This approach yields a total of 10152 videos with modified audios, which is approximately 8% of the total training set. Of these, 4209 have unchanged sample rates but slightly difference samples, while 5249 have different rates from the original. Within the unchanged sample rate, one almost perceive no difference by listening to it, for these clips the audio difference could be result of transcoding, dynamic compression or simply noise removal, on the other hand, the difference is much more pronounced when examine videos with large difference in sample rate, obvious distortion or fluctuation in pitches can be easily identified with human ear.

There are multiple steps of transformation that’s required in order to assess the difference in the audio, to start we utilized the FFmpeg function to extract the wav file from the original video, this is a open source project that consists of number of libraries that handling video, audio and other multimedia files and streams. Once the wave files are generated, we can convert them into spectrograms, which is a visual representation of the signal frequency as it varies with time, this allow us to create “voiceprint” of the wav file and convert the image into an array for arithmetic calculation, and it is through vector calculation based on these spectrograms we are able to determine sample rate differences between the fake and the original videos. Below is a side by side comparison of spectrograms for the video “asmpfjfzif” and “dzyuwjkjui”, while they are not easily distinguishable visually, their np array in sample rate difference is over 4000 and the audio is clearly altered compare to the original.





Knowing which video is altered in image or audio allow us to take a ensemble approach where we can train different models based on the nature of the modification. And we certainly would like to explore this strategy going forward. However, given the constraint on time and computation power this is not feasible for our current experiment, as the primary workhorse is still the image classification model. Fortunately, in all of 10152 altered videos there is at least minimal change in the pixel data, in other words no deepfake videos exist where the audio is fake and the face is real for all frames, therefore obviate the need to treat those videos separately.