I. INTRODUCTION

This project's goal is annotating and analyzing audio recorded in everyday environments, with a focus on the content regarding sounds that can be heard at the scene. We start the project with a set of pre-recorded audio data. Audio files provided for this project are a subset of TAU Urban Acoustic Scenes 2019, 150 files. Data was recorded in 12 different large European cities during 2018-2019, in scenes including parks, streets, trams, shopping malls... Data is provided as individual clips of 10-seconds length and is openly available and contains information on acoustic scene, city, and location ID (location ID indicates different parks, streets...). There are 3 parts in this project: annotation of everyday audio, audio analysis and lastly, writing report.

II. DATA ANNOTATION PROCESS

1. Describe the annotation process

Data annotation is the process of giving metadata: tags, or labels, short description to the data we observe, or specifically in this audio processing project, listen to. In this project, we used a website to do this annotation task. The process include hearing, select available tags corresponding to sound heard in the audio, and finally a short description of the audio. All audios heard in annotation tool do not have any information to avoid bias when annotating.

There are few good and bad things about this annotation process from the annotator's point of view. Regarding bright side of the task, this task is one of the most important tasks of the entire project since without it, or without doing it carefully, we can go far anywhere. For example, if we want to build a machine learning model to learn to classify sound of the audio, training the model with many false labelled data will not help the model learning much. Therefore, the annotator has a huge impact to the project. Moreover, if the annotator happens to be also the person to do other tasks afterwards, such as pre-processing, analyzing data, he/she have a good idea of what to do with the data, and build a good pipeline for those tasks. However, the amount of data is not small, usually hundreds, or sometimes, thousands of them, and the process is quite repetitive and sometime daunting. Particularly with this project, the audios are raw, and they contain a lot of sounds mixing together that are hard to differentiate. The annotator can consider those sounds as noise and move on. This is fine, but given the pressure of quality labelled data, he/she must try a little bit more and identify it. Overall, the task itself is not hard, in terms of technical skills, but it is long, 4 to 5 hours of just listening audio and give them labels.

As discussed above, the task, in general, can be found as boring, and this could decrease annotator's efficiency in the task. Therefore, some suggestions and improvement could be considered to make the process easier, and more enjoyable. First of all, splitting the whole task into smaller sessions, as suggested in the instruction, is definitely a good idea. Secondly, the annotation tool is too simple, and thus, make the task itself boring when doing it for long hours. Developers can use gamification strategy when designing this annotation tool. In other words, some game-design elements can be implemented into the tool. Thirdly, the option of making an application for the tool could be considered. Until this point, the web design of the tool is not optimized for small-screen devices like smartphone, but we know that people usually carry their mobile devices more than computers, or tablets when being outside. Finally, besides a simple progression bar on top of the page, maybe the annotators should know how many they have to do left, and estimated time left for the task. By doing so, not only they get motivated by getting closer and closer to the finish line, but they could also estimate their current workload in each session.

2. Dataset statistics

After the annotation task, we were given a subset of what we have annotated, 131/150 files to do the analysis part, and below are some statistics collected from what we were given.

Data statistics			Values
Number of files			131
Number of labels			10
Numbers of files annotated with	0	classes	**8**
	1		66
	2		45
	3		12
Numbers of audio files annotated as	adults talking		**79**
	birds singing		33
	announcement speech		2
	children voices		29
	footsteps		29
	traffic noise		9
	dog barking		2
	siren		4
	music		3
	announcement jingle		2

Table 1: *Table of statistic*

As we can see from the table above, there is a total 131 audio files in the dataset but 8 of them are not annotated (because there are no suitable available tags to annotate the audio, a short description is written instead, but the description is not in the scope to be analyzed in this project). Therefore, we end up with 123 audio files to be analyzed with most of them have 1 or 2 tags. Among all files, we can see that "adults talking" present in the majority of files. If we take a closer look at the column "fileName" in the annotation .csv file (open the file with Excel app, sort the data by this column), there are 3 kinds of locations where audios are recorded: airport, park, and public square. And as we look more closely to this sorted data, we can see a pattern of annotation that help us understand the tag "adults talking" is labeled the most. As shown in the **figure 1** below (this snippet is only an example, but one can open the file and verify the result), almost all recording locations at the airports or public squares are labelled with "adults talking". This may indicate that audios recorded in these areas are crowed and noisy and could also contain some other mixture classes. Nonetheless, we can also see a pattern where this label rarely exists in audios recorded at the parks, and instead, "birds singing" exists most often here. In short, after a quick overview of data, in the next section, when calculating similarities between each file, we can guess that two-thirds total number of files quite similar to each other.

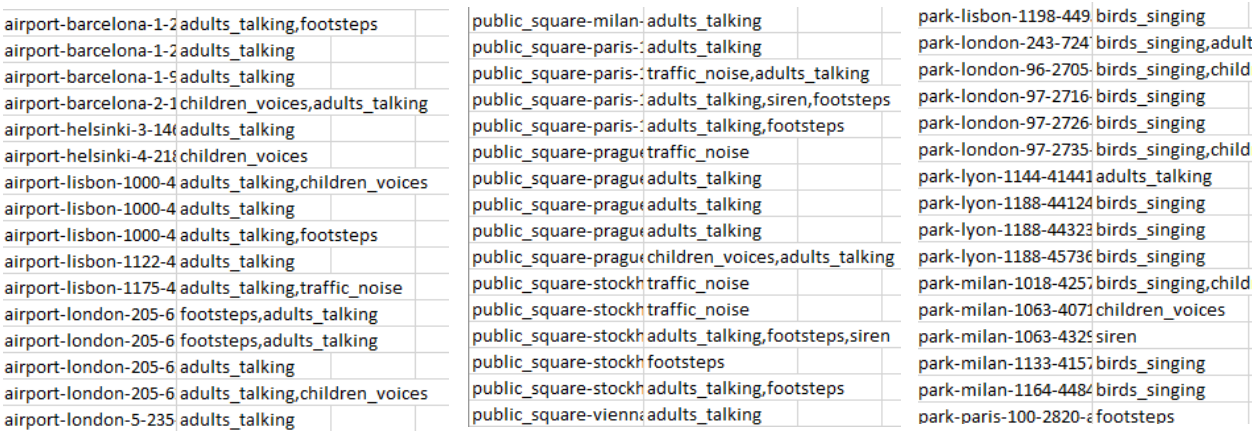


Figure 1: *Example of snippets from annotation file*

III. AUDIO ANALYSIS

1. Implementation

In general, with speech processing, before we even do any analysis, one should always try to represent the original audio files in some meaningful way to detect audio's features and mimic human ear's perception of sound. Mel-frequency cepstrum coefficients (MFCCs) is the most popular method for various reasons, and one of those is the ability to represent the spectral energy distribution in a perceptually meaningful way known. It is commonly used to various tasks like speech recognition, or audio segmentation, and so on. Here, we use a function called librosa.feature.mfcc (from a Python library called librosa) to get MFCCs of audio files to do our analysis task. Details how to program this can be found in the file main.ipynb. Besides common arguments passed into this function like audio signal *y* and sampling rate of audio *sr*, we also pass some additional arguments: *n_mfcc=40* (equivalent to around 0-8kHz, just a common choice), *norm='ortho'* (defeault behaviour), with some additional keyword arguments to function librosa.feature.melspectrogram: *win_length=20ms* (we are dealing with speech audio, so we should use short window length instead of long one like 40ms), *n_fft=win_length* (default behaviour), *window=hamming(win_len, sym=False)* (better visibility of spectral peaks and reduce spectral smearing, compared to rectangular windowing/no windowing, *sym=False* since we are doing spectral analysis), *hop_len=win_len/2* (Common hop size in windowing process, 50% overlap-added).

We then need to compute the similarity between each file by: computing mean and standard deviation of 40 MFCCs over the temporal axis and stack them together (result in 80x1 vector for each file), and then compute Cosine similarity on every pair of files. The range of value is from 0 to 1with 1 means 2 files are very similar and 0 means 2 files are very different.

2. Results and discussion

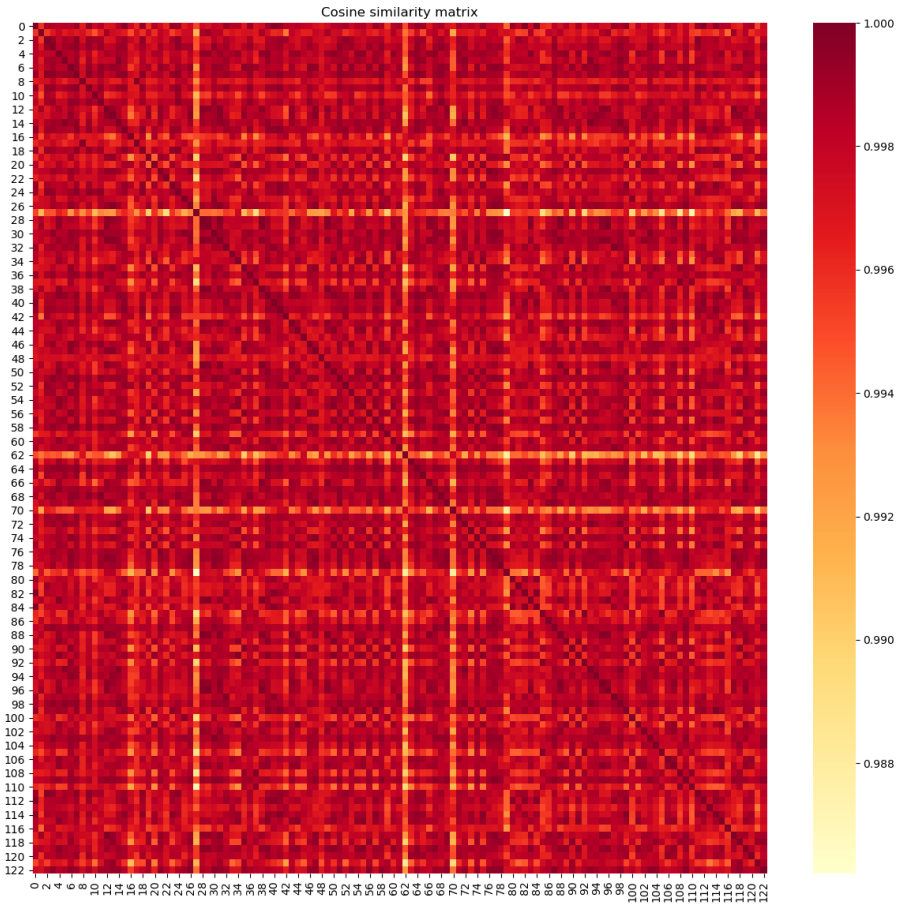
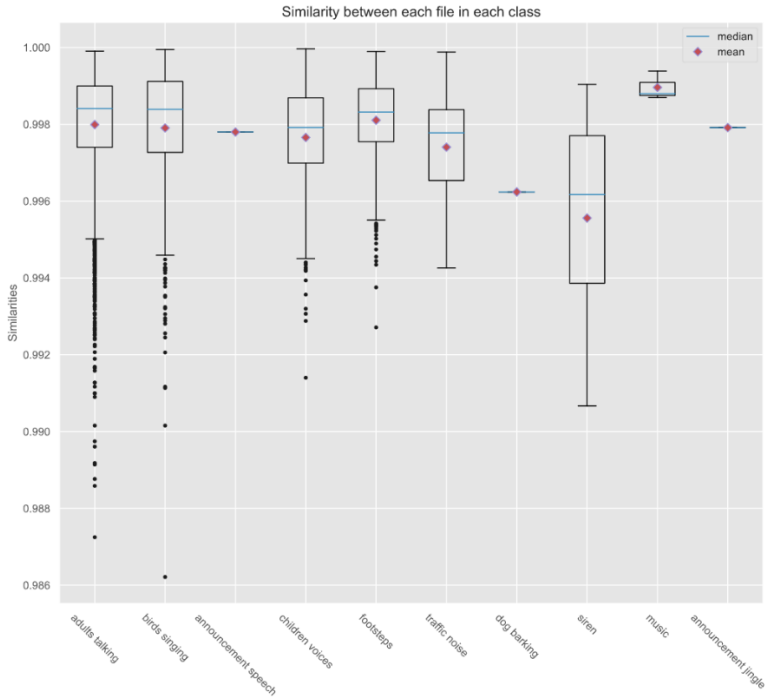


Figure 3: *Heatmap of similarity between each file*

In the previous part, we guess that two-thirds total number of files are quite similar to each other. However, now, by looking at the heatmap, in general, we can see that audios are quite similar to each other. This might indicate that most audios are recorded at noisy place, and places having bird singing, without adults talking may have children voices (still humans) and the frequencies of humans' voices, in general, are not too different, compared to birds sounds.



Class	Average similarity
adults talking	0.9979921
birds singing	0.9979054
announcement speech	0.9978002
children voices	0.99765843
footsteps	0.9981074
traffic noise	0.997406
dog barking	0.99623924
siren	0.99556035
music	0.9989617
announcement jingle	0.99791485
Overall	0.98958944

Figure 4 and table 2: *Similarity between each file in each class, and average similarity of each class*

If we take a deeper look into similarity between files in the same class, shown in the **figure 4** and **table 2** above, we have some interesting insight. Dog barking has the least average similarity. There are only 2 files having this tag and these files are recording at the airport and park. This may mean that sound from these scenes is pretty different from each other. The same situation may also happen with siren which 4 files are recorded at 3 different places. However, music and announcement jingles, even though each only has 3, and 2 files respectively, but have very high average similarities. Hence, we could say that files with these labels are recorded at the same scenes. Regarding set of classes that usually go together, we could see those classes are adults talking, birds singing, children voices, footsteps, and maybe even traffic noise (hard to conclude since there are only 9 files). They all have the average similarity around 0.997 to 0.998 and the number of files having these labels are from 29 up to 79. The locations recording these audios seems to be very noisy with lots of humans, and they can be open spaces.

IV. CONCLUSION

This audio analysis project has given us an overview of the daily life sounds. We first annotate different audio files and then analyze, discuss the result, in particular, about the similarity of audios. Typically, this kind project can be further expanded with some tasks like audio segmentation or audio recognition, but maybe with larger dataset.