3. Audio Analysis

3.1. Implementation

To be honest, I almost leave all parameters as default value such as:

+ n\_fft = 2048

+ hop\_length = 512

+ n\_mels = 128

I believe it has trade-off between accuracy and program size/memory cost and calculation time. So default value may give me balanced between 2 factors.

The parameter that I tried to change was n\_mfcc = 40 and 80. However, the results does not vary much.

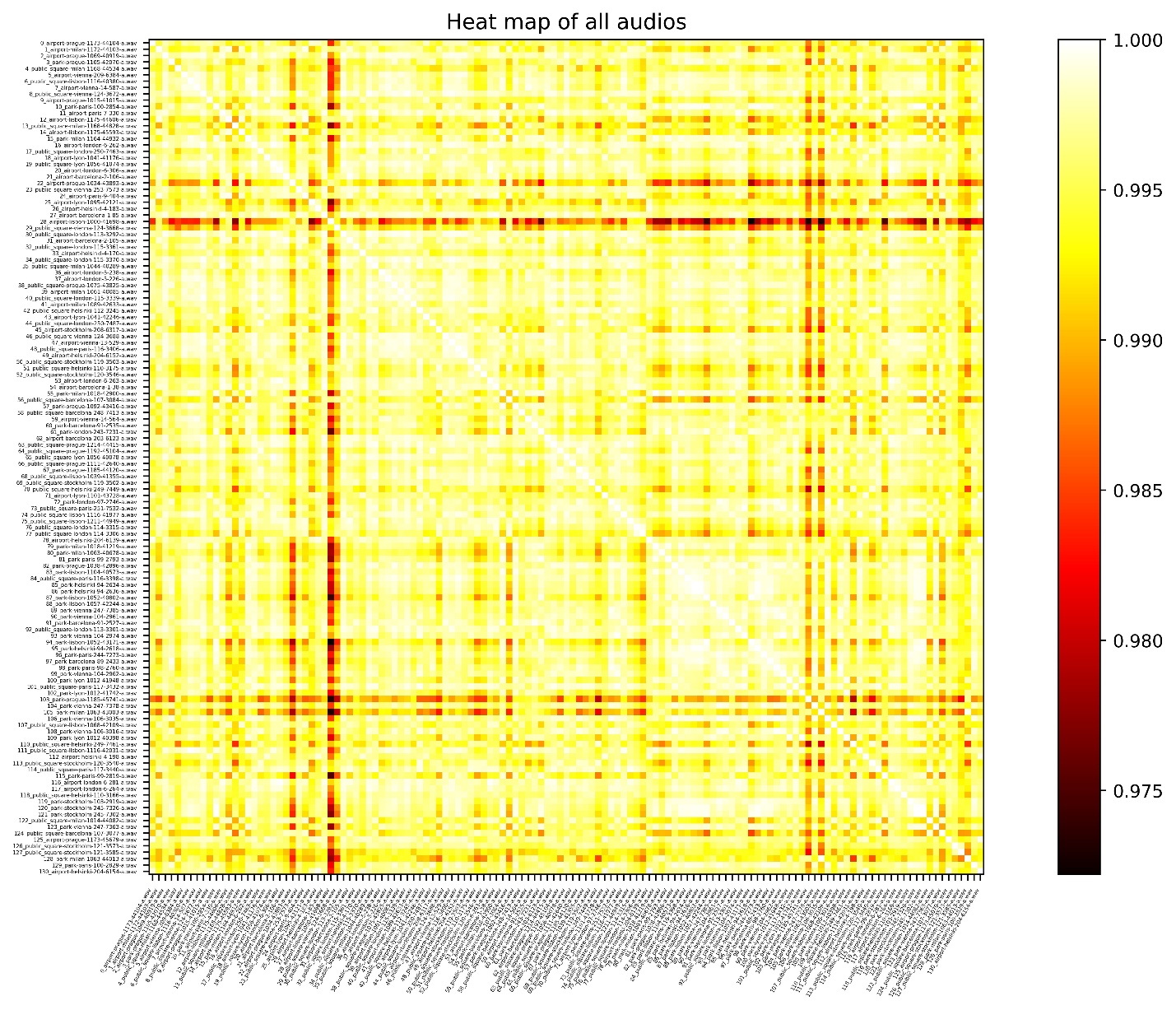
I implemented both methods including cosine similarity and DTW distance. However, in this report, I would mostly focus on Cosine Similarity.

3.2. Results and discussion

3.2.1. Heatmap and average similarity

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Class | 1 | | | | | 2 |
| Sub-class | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 2.1 |
| Description | Adults talking | Adults talking+  children voices | Adults talking+  footsteps | Adults talking+  Birds singing | Adults talking+  Traffic noise | Children voices |
| Average  similarity | 0.99664 | 0.99544 | 0.99751 | 0.99664 | 0.99715 | 0.99825 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Class | 3 | | 4 | | 5 | | 6 | 7 |
| Sub-class | 3.1 | 3.2 | 4.1 | 4.2 | 5.1 | 5.2 | 6.1 | 7.1 |
| Description | Birds singing | Birds singing +  Traffic noise/ footsteps | Traffic noise | Traffic noise +  footsteps | Footsteps | Footsteps  +  Other sound | Siren  + other sound | Unclear-ed |
| Average  similarity | 0.99853 | 0.99772 | 0.99454 | 0.99695 | 0.99744 | 0.99949 | 0.99510 | 0.99600 |



3.2.2. Report from result data

- By arranging audios from same class nearby each other, heat map graph has fairly bright pixels (high level of similarity) along the white sanity check line.

- I can observe that within the same class, the first sub-class, which I marked with only one pure sound, often has higher level of similarity compared to other sub-classes.

- The results vary little (from 0.975 to 1). However, class 1, “adults talking” similarity levels wildly fluctuate. In my opinion, people voices are more diverse than other sound such as birds, footsteps or siren, which are more typical.

- I have also tried with DTW distance and got the same result (the levels are inverted because distance and similarity are invert of each other).

4. Conclusion

From my point of view, the project is interesting, despite the cost of time and effort. It provides practical knowledge and training to process real sound. I am aware of difficulties when analyzing these real data that it can cause confusing. To sum up, through this project, I have learnt the purposes of different analyzing methods, practiced using them and what information can be retrieved working with real audio.