Environmental audio analysis

COMP.SGN.120-2020-2021-1 Introduction to Audio Processing
Project work

Aliisa Nissinen
Student ID: 291603
13.12.2020

1. Introduction

This project work was a mandatory part of the Introduction to Audio Processing course. The idea was analyze 131 different 10 seconds audios and how they were similar to each other.

Audio files provided for the work are a subset of TAU Urban Acoustic Scenes 2019. Data was recorded in 12 different large European cities during 2018-2019, in scenes including e.g. parks, streets, trams, shopping malls.

This project work included the annotation of the data and the data analysis. The conclusions were made based on the results of the analysis.

Before the analysis, the assumptions were that the audios recorded in the same environment or with the same voice tags are more similar to each other than others.

2. Data annotation process

The first part of the project was data annotation, where all given audios were tagged with suitable voice tags, including e.g. footsteps, adults talking, birds singing. Also, for each audio had to be written a minimum of 5 words definition.

2.1. Describe the annotation process

The annotated data were 10 seconds length audios, that have been recorded at streets, parks, trams, etc. It was possible to listen to those audios as many times as you wanted. The best way to hear all voices in those audios was with anti-noise headphones, that was used in this project work.

Most of the data was the same type and it was very hard to concentrate on them. Especially the 5 words definition was hard to come up with, because some of the audios had only the traffic noise and some footsteps in the background. However, some of the audios had unexpected voices like sirens or a dog barking that brought a little variation to the annotation process.

The entire data annotation part took about 3 hours. It would have been faster if the 5-word definition hadn't been mandatory and you could only write it if you came up with something special.

2.1. Dataset statistics

The annotated data had 131 audio clips, 41 of them have been recorded at the park, 36 at the airport, and 54 in the public square. Audios have been recorded in different countries like Milan, Paris, Lisbon, and Stockholm. With the annotation process, all audios were tagged with given tags and, also own tags could be created. Some of the audios don't have any tags if there wasn't any recognizable sounds. Below are statistics about the given tags.

Tag	**Audios**
footsteps	69
traffic_noise	52
adults_talking	94
children_voices	26
siren	2
announcement_speech	7
music	15
announcement_jingle	1
birds_singing	23
dog_barking	2

Two of the audios didn't have any tags. On average, each soundtrack has approximately 2 tags, the variation was between 0 and 4.

Based on this most of the audios had footsteps and adults talking, which can also be deduced from the recording locations. Also, it is obvious that if 2 of the locations are outdoors, then the announcement speeches and jingles are not very likely in those audios.

In the analysis is one part, where analyzed how soundtracks from the same category are similar to each other. In that part analyzed only the categories that have more than 2 audios, so siren, announcement jingle, and a dog barking categories weren't taken into account. It's because it's hard to make a conclusion for such a small amount of samples.

3. Audio analysis

The second part of the project was audio analysis, where audios' similarity was analyzed. The similarity was calculated of each file to each file, also was calculated an average similarity between the same categories and between all the data.

3.1. Implementation

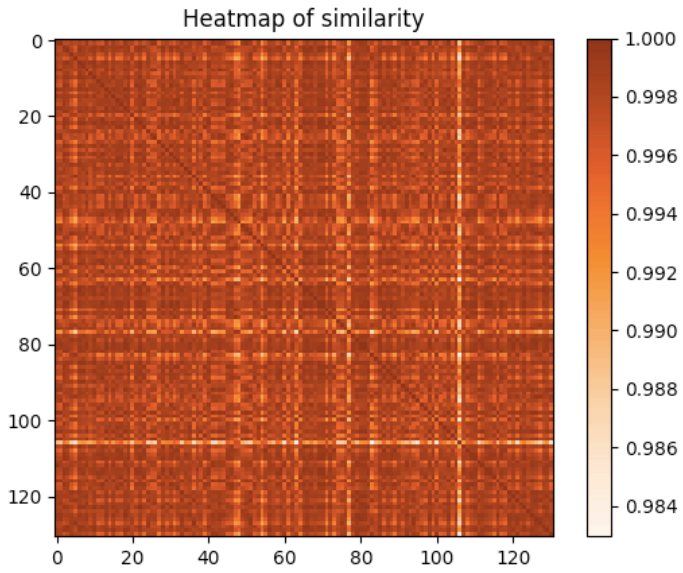
First was implemented matrix that contains cosine similarity of each file to each file. In every row was a similarity of one audio to all other soundtracks. So in the diagonal axis there are only ones, because the similarity of audio to itself is one. In this part first was calculated the MFCCs for each file, with 40 frames that were given in the exercise instructions and with FFT window size 512 that was used in most of the exercises in this course. After that was calculated the features of the MFCC, the standard deviation and the mean value, which were put in one vector, size of 80x1. In the end, calculate the cosine similarity matrix, size of 131x131, that is the same as the number of audios in the data. The whole calculation took about 3 hours for my old MacBook laptop so the matrix was also written in the cvs file, so it was easier to continue the project work, without running the whole process over and over again.

After the cosine similarity matrix, was calculated the average similarity for each class. Different classes were introduced in chapter 2. This was implemented with the cosine similarity matrix, by removing samples not belonging to a specific category.

In the last part was calculated the average similarity between files for all data. Again I used the cosine similarity matrix, from where I removed the diagonal values.

3.2. Results and discussion

Below is the heat map of the cosine similarity matrix. From the heat map can be seen, that most of the data are very similar. The diagonal dark line is a similarity between files to itself, that is one.



The color in audios number 106 and 77 are lighter than in others. Those audios were recorded in the park and public square, and they have tags traffic_noise, footsteps, adults_talking and birds_singing. Most of the audios have these same tags, so can't make a conclusion from that. The difference in similarity values are very small, all result is between 0.98-1.00.

The average class similarity results are presented in the table below.

Tag	**Average similarity**
footsteps	0,998098
traffic_noise	0,998112
adults_talking	0,997974
children_voices	0,997681
announcement_speech	0,997845
music	0,998288
birds_singing	0,998328

Again the difference in the values is very small. The reason can be, that most of the data contain the footsteps and adults talking and other sounds weren't as much present in the audios. Also, the audios were very quiet and all of them had a similar hum in the background. The children voice class is least similar, the reason can be that most of the children's voices were laughing and screaming that was different from other sounds and a little bit louder.

The average similarity between files for all data was 0,997686, which is very similar to class similarities.

4. Conclusions

The assumption, that audios that were recorded in the same environment are more similar, don't show in these results. Also, different categories didn't show a clear difference in similarity. In this project, the result was, that all files were pretty similar and the results didn't show any interesting information.

The reason for these results can be in audios because they were so short and that's why they didn't have much difference, also the quality was bad. The voices were heard only with the anti-noise headphones. The differences could have come out better perhaps with more advanced computing methods.