Tampere University

COMP.SGN.120
Introduction to Audio Processing

Project Work Description

17 Nov 2020

Annamaria Mesaros

Project: Environmental audio analysis

The project aims to collect and analyze audio recorded in everyday environments, with a focus on the content in terms of sounds that can be heard at the scene.

We start with a set of already recorded audio data.

- 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.
- Data is provided as individual clips of 10-seconds length
- Data is openly available and contains information on acoustic scene, city, and location ID (location ID indicates different parks, streets, etc)
- For details, see http://dcase.community/challenge2019/task-acoustic-scene-classification#audio-dataset

We started by annotating sounds that can be heard in each audio file, given a predefined list.

Project steps

- Part 1: Annotation of everyday audio. (3 p)
- ~~Part 2: Data curation (1.5 p)~~ points have been transferred to final report
- Part 2: Audio analysis (3 p)
- Part 3: Final report (4 p)

Due to objections to the amount of work, data curation part is left out and its points are moved to the final report. Make sure to write a proper report!

Submission in Moodle by 17 Dec 2020, no extension!
Submission consists of code (py or ipynb) AND pdf report with a given structure.
More details in the later slides.

Part 1: Annotation of everyday audio. (3 p)

Each assignment contains about 130 files to be annotated. Deadline: 19 Nov (Thursday)

Evaluated based on person activity statistics in the annotation tool.

Part 2: Audio analysis

Research question:

what is the average similarity of two files labeled with the same sound (e.g. footsteps)?

Implementation:

Compare average intra-class similarity between the classes and overall average similarity of all data. Discuss acoustic content based on your observations.

Starting point: You have a collection of clips annotated with X classes. Some clips belong to multiple classes because there are multiple sounds present in them. You will analyze how prominent the characteristics of each class are in the data. (only based on the sound labels, not using the description sentence)

Each student will receive the set of files and the corresponding annotations they produced. The data subset and associated annotations will influence the analysis outcome and discussion.

Do not modify the information you receive! (sound classes, etc)

Part 2: Audio analysis (2)

You will be given:

- a zip file containing all audio clips that were annotated collaboratively
- a csv file containing a list of files annotated by you
- We are working on generating these and will give them as soon as possible. (You have the MFCC exercise next week anyway so that's a good start.)

You annotated a small random subset of the available files; you will analyze your own annotations

The csv file will be named as your student number

- Please use in your analysis the exact list of clips provided in your personal csv file
- You will return the code, make sure the part where you read the required input file list is visible/indicated

Step 1: Calculate a similarity matrix between the files

Using cosine similarity between aggregate feature vectors

- Calculate MFCCs for each file (for example 40 filters x t frames)
- Aggregate the calculated features into a single vector over time as mean and std (so with 40 filters you obtain an 80x1 vector, with 40 values being average and 40 being std of the corresponding order coefficient; you calculate mean and std of the 40 coefficients and stack them)
- Calculate cosine similarity of each file to each file as matrix S; matrix S has elements s(i,j) which represent cosine similarity between file i and file j
- Sanity check: similarity of a file to itself should be 1!

Note: you can use librosa for the MFCC, no need to implement your own.

Step 1: ADVANCED version

Alternatively, instead of cosine similarity you can use dynamic time warping to calculate a distance matrix (DTW comes in the lecture in 26 nov)

- Calculate MFCC for each file (for example 40 filters x t frames)
- Calculate DTW cost for aligning each file to each file and represent as a matrix D of distance between files, having elements d(i,j) as distance between file i and file j
- DTW comes in lecture 11
- Sanity check: cost of aligning a file to itself should be 0.

Note: you can use librosa for DTW.

Use either one of the two methods, it's up to you.

Make sure to analyze the results accordingly.

Step 2: Calculate average similarity

Calculate average similarity between files for each class

- Extract from S all elements s(i,j) for which files i and j are tagged with the class you are processing at this time; exclude the diagonal elements
- Calculate average similarity for the analyzed class; repeat for all classes
- In case of using DTW, what you get is average distance for each class.

Notes:

- In some classes there will be the same files, since there are multiple sound labels for most files. This is of course normal.
- What you want to see here is if the characteristics of one class are more prominent than others.
- If you use DTW, pay attention to the difference between similarity and distance (cost)!

Step 3: Average similarity of all data

Calculate average similarity between files for all data

- Calculate average of the entire matrix (similarity or distance, depending on the method)

Step 4: Analyze the content of your data

Analyze the data content based on the obtained values. Think about the following:

- What do the obtained values indicate? Which class is most similar in terms of acoustic content? What is the difference, conceptually, between similarity and distance?
- How similar is the content of individual classes compared to the entire dataset? What does this mean?
- What about sets of classes? Can you find groups of sound classes that occur together often and create a very similar environment? For example footsteps and dog barking, etc, that are more prominent as a set than each individual class. Do they have anything else in common? (e.g. same acoustic scene, same location? - is this something that explains more the similarity than the individual sound class?)

  Hint: look at the file-based values and analyze co-occurence of tags for the files that are most similar. You can do this subjectively (without calculating, just observing values and discussing them)

Part 3: Final report

Write a report with the following structure:

(maximum 4 pages, font size 12, single spacing, 2 cm borders)

1. Introduction

   Here you describe the audio analysis task and assumptions it is based on

2. Data annotation process

   2.1. Describe the annotation process

   What was annotated?

   What was good and what was bad from annotator point of view?

   How would you improve/modify the annotation task to make it simpler/faster?

Part 3: Final report (cont.)

2.2. Dataset statistics

Calculate some statistics of the dataset you received for analysis (e.g. how many classes of sounds are present, which are the most/least frequent, how many classes on average per clip, etc). **Add to the report tables** with such information and discuss content from this point of view.

3. Audio analysis

3.1. Implementation

Explain the method you implemented; include details such as choice of parameters for calculating the features (window size, number of filters, fft size, with motivation of the choice).

3.2 Results and discussion

Analyze the results following the structure given as questions from Part 2 Step 4. Include both numerical analysis and subjective analysis/intuition and interpretation of what is going on in the data.

4. Conclusions

Short conclusion about the entire process

Project return

In Moodle, by **17 Dec 2020**. No exceptions and no extensions. Return:

- **Code** as py or ipynb, make sure to comment it to at least indicate main steps
  - **Code is worth 3 points as data analysis.**
  - Grading will be based on correctness and comments

- **pdf report** with given structure and content. Have a title (Environmental audio analysis) and author name.
  - **Report is worth 4 points.**
  - Grading will be based on structure, method presentation and the quality of the results analysis and discussion.
  - Include in the report numerical results and their interpretation, possibly figures, **heatmap plot of similarity** (distance) matrix, **table with average values** per class.

Few more practical points

- You can use library functions for the MFCC or DTW, no need to implement it yourself
- There is no restriction on use of libraries and functions (e.g manipulating csv files)
- There is no “very wrong” thing to write in the report. Please analyze the data you have, in terms of what happens, how sounds are related or differ from each other, how they appear together, give your opinion on why certain things/phenomena are observed. Same goes for the data annotation process analysis.
- Support in slack for project will start after exercise 5 deadline (because ex 5 is MFCC). We will try to have later one dedicated live slack session for project help.