MLPC 2024 Task 2: Data Exploration

Katharina Hoedt, Verena Praher, Paul Primus, Florian Schmid

Institute of Computational Perception
Johannes Kepler University Linz
March 18, 2024

Context

Our overall objective for this year's project is to develop a system that can detect speech commands in audio recordings. The purpose of this system is to allow users to control different devices in a smart home using speech commands. Read the *Project Description* on the Moodle page for a detailed project outline.

Developing such a keyword recognition system with machine learning entails the following steps:

1. record a training set of keywords and non-keyword sounds
2. compute a set of candidate audio features
3. perform a thorough analysis of the features and select useful features
4. train and evaluate a range of classifiers and hyperparameters; find the ones that can distinguish between keywords and unrelated sounds, such as "Other Speech/Noise" and "Silence"
5. apply and evaluate the trained classifiers for detecting speech commands in everyday scenarios and select the model that works the best

After the collective effort of Step 1, we computed a variety of audio features for you. Now we reach Step 3 in this process. Here we want you to work in small teams and investigate the data/features thoroughly.

Task Outline (22 + 3 points)

Deadline: April 10

Instead of blindly throwing that data at machine learning algorithms, a good data scientist will first do some exploratory data analysis. In your team, we would like you to analyze the data and write a report answering questions about a few different aspects. Furthermore, you will have to prepare a few slides, summarizing your findings about one specific topic.

Participating in Task 2 is a requirement to pass the course.

Written Report (max. 22 points)

For the first part of this assignment, you will have to write a report based on the **template** provided to you on Moodle (item *MLPC Report Template*). In this report, you will address the following questions:

1. Data Consistency & Quality (5 pts)
    a. Without listening to all recordings (i.e., by examining feature statistics): Are there any obvious inconsistencies (wrongly labeled audio snippets, outliers, etc.) in the data?
    b. Listen to some of the recorded words. Are there notable biases in the data set? Are the collected samples representative of speech commands that will be encountered once the application has been deployed?
    c. Listen to some of the sounds labeled as “Other”. Which types of sounds can you identify? Are they representative of everyday sounds in a household?
2. Label Characteristics (5 pts)
    a. Concerning the upcoming task of identifying speech commands in everyday scenes: How would you group the 20 words and audio snippets labeled as “Other” in classes and why?
    b. Are the resulting classes unbalanced? If so, how much?
3. Feature Characteristics (5 pts)
    a. How are the pre-computed audio features distributed?
    b. Are there any pairs or subsets of features that seem highly correlated or redundant?
    c. Are there differences in the feature distributions for different speakers?
4. Feature / Label Agreement (5 pts)
    a. Which features seem useful for classification? Which ones are correlated with the labels?
    b. Do similar words (e.g. Haus - aus, Ofen - offen, or nicht - Licht) have a similar feature distribution?
5. Conclusion (2 pts)
    a. Which conclusions can you draw from your analysis?

In addition to addressing these questions, you will also have to add a **statement of the contributions** of all team members as indicated in the template. The report should not exceed a **page limit of 5 pages**, of which in total **at most 3 pages** should be **text**.

Slide Set (3 points)

For the complementary slide set, you will have to try to present some of the results of your written report in a clear and concise manner, such that you could show the result to fellow course participants. More precisely, you will have to answer **all questions of one sub-topic** introduced in the previous section. The specific topic is determined based on the **first letter of your group-name**, i.e. A for Team Aberrant, or B for Team Bed. To find your topic, determine the according letter, and find your topic in the following list:

First letter of group-name	Topic
A, C, E, M, Q	1. Data Consistency & Quality
B, F, I, L, N, P	2. Label Characteristics
D, G, J, R, T, U, W	3. Feature Characteristics
H, K, O, S, V, Y, Z	4. Feature / Label Agreement

The **upper limit for the number of slides** you should prepare is **4** (excluding an additional title slide that should contain your group name and the member names).

Dataset

The dataset download links are available on Moodle (by March 19), and the format and content of the dataset are described in detail in the slide deck for Meeting 2 (March 18). Please refer to that slide deck for information on the audio features and the file formats.

Grading

The written reports for this task are evaluated according to the following criteria, for each of the five topics given in the task outline:

- **Thoroughness & Completeness**: Have you thought about the problem, and answered every question?
- **Clarity**: Are the ideas, features, algorithms, and results described clearly? Based on your descriptions, could the reader reconstruct your experiments?
- **Presentation**: Did you select an appropriate way of communicating your results, e.g., did you use meaningful plots where helpful?
- **Correctness**: Is the proposed procedure/experiment sound, correct?
- **Punctuality**: The reports must be submitted in time. Any delay will result in reduced grades. Specifically, submitting on April 11 will deduct 1/3 of the points, submitting on April 12 will deduct 2/3 of the points, and submissions on April 13 or later will be rejected.

For the slide set, you will be awarded points if you have a valid set (i.e. within the slide limit) submitted for the assigned topic.

Summary

- **Completing Task 2 is a requirement to pass this course.**
- Look at the given questions, and answer **all** of them appropriately in a written report. Make sure to use the **report template** provided to you via Moodle. Adhere to the given **page limit** (max. 5 pages where at most 3 pages can be text) and include a statement about the contributions of all team members.
- Create a set of slides tackling the questions of **one** of the topics. The topic is determined by the first letter of your group name. Make sure to adhere to the **slide limit** for this step as well (max. 4 + 1 title slide).

- Upload the written report as well as your slides to Moodle (*Upload Step 3: Submit your team's Task 2 reports*) until **April 10th**.
- You will get a maximum number of 22 points for your written report, and 3 points for the slide set.