MLPC 2024 Task 3: Classification

Katharina Hoedt, Verena Praher, Paul Primus, Florian Schmid

Institute of Computational Perception
Johannes Kepler University Linz
April 15, 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 speech command 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 their corresponding hyperparameters; find the ones that can distinguish between keywords and unrelated sounds
5. apply and evaluate the trained classifiers for detecting speech commands in everyday scenarios and select the model that works the best.

After joining our forces to collect a data set of appropriate size, we computed a variety of audio features for you (Steps 1 & 2). In the previous assignment (Step 3), you then conducted a thorough analysis of the data set and the features. We've now reached step 4, where we finally get to train some classifiers! As for the previous assignment, you will work on this task with your team. Your two main objectives will be (1) to design an evaluation setup and (2) to develop a machine-learning pipeline for training a variety of classifiers, logging/visualizing the results, and selecting the best models.

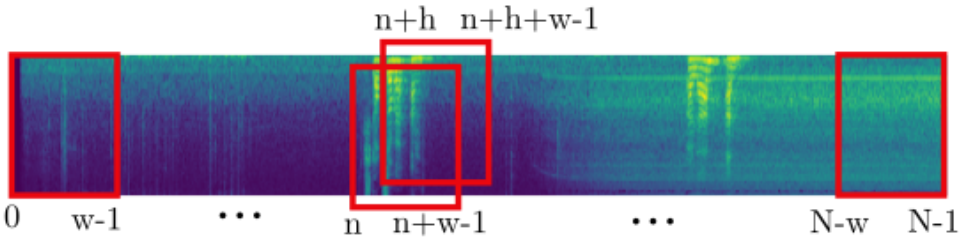
Task Outline (37 + 3 points)

Deadline: May 22

Perform systematic classification experiments in your team using the same training data as for Task 2:

- Focus on predicting the correct class (keywords and other classes) for each 1.1-second snippet.
- Decide on an evaluation criterion to use. Mind that your goal is to recognize all keywords correctly and to distinguish them from other sounds.

- Use a proper data set split for selecting the best classifier and hyperparameters and to assess the final performance. Based on what you know about the data, decide how to split it into folds. Avoid information/data leakage across folds. Remember that your goal is to estimate how your classifier will perform for unheard speakers.
- Apply at least three different learning algorithms from different major groups: Support Vector Machines, Neural Networks, Nearest Neighbor Classifiers, Naive Bayes, Decision Trees, Generalized Linear Models, Linear and Quadratic Discriminant Analysis, etc.; maybe also an “ensemble method” such as a Random Forest.
- For each algorithm, systematically evaluate different hyperparameter settings, especially those that control the algorithm’s overfitting behavior. Analyse and document how the hyperparameters affect whether overfitting occurs (and to what extent it occurs) and how they affect classification performance.
- The last step is preparation for the upcoming challenge phase: Qualitatively evaluate the performance of your best classifier on the 4 scenes provided via Moodle. Those scenes are 6-24 seconds long and contain complete speech commands consisting of a device keyword and an action keyword (e.g., “Radio aus”). To apply the classifier from the previous stage to these longer recordings, they first need to be cut into shorter 1.1-second segments. To this end, use a sliding window with a hop size of 1 frame to extract feature sequences of 44 frames (44 sequence steps correspond to 1.1 seconds). The figure below illustrates the process of extracting snippets (red windows) from a mel-spectrogram of length N with a hop size of h frames and a windows size of w.



For these experiments, you may use any machine learning package you want, such as scikit-learn for Python or built-in toolboxes for Matlab. You may want to exploit that neither all the features nor all the training examples are equally useful. Especially when working on (not-so-powerful) laptops, it can pay off to subsample the training data or select a set of non-redundant features to save computation time.

Participating in Task 3 is a requirement to pass the course.

Written Report (max. 37 points)

For the first part of the submission, 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 Split**
    a. Describe the data split you used for model selection and hyperparameter tuning and the reason for your choice.
    b. How did you avoid information leakage between the sets?
    c. How did you derive final, unbiased performance estimates?
2. **Classes & Features**
    a. Briefly describe how you grouped the 20 words and audio snippets labeled as "Other" into classes and why.
    b. Which subset of features was selected? How did you select the subset?
    c. Did you apply preprocessing, and if so, which and why?
3. **Evaluation**
    a. Which evaluation criterion did you choose to compare hyperparameter settings and algorithms, and why?
    b. What is the baseline performance? What could be the best possible performance?
4. **Experiments**: For at least **three different classifiers** from the major groups given above, systematically vary the most important hyperparameters and answer the following questions for each of them:
    a. How does classification performance change with varying hyperparameter values? Visualize the change in performance.
    b. (To what extent) Does overfitting or underfitting occur? and what does it depend on?
    c. After selecting appropriate hyperparameters, compare the final, unbiased performance estimate of the three classifiers.
5. **Analysis of Realistic Scenes:** Qualitatively evaluate your best classifier on the 4 scenes available in Moodle.
    a. Listen to the scenes and inspect the corresponding predictions of the classifier. How well does the classifier recognize keywords?
    b. What are particular problematic conditions that cause the classifier to miss or mispredict keywords? How could you alleviate this problem in the challenge phase of the project?
    c. Use the spectrogram and the sequence of predictions to visualize interesting (positive or negative) findings.

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 must not exceed 7 pages, of which at most 4 should be **text**.

Slide Set (3 points)

The second part of the submission is a short presentation. This complementary slide set serves to present selected results in a clear and concise manner to your 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, B, C, L	Data Split
D, E, F, G	Classes & Features
H, I, J, N, Y	Evaluation
K, M, O, P, T, Z	Experiments (one classifier type)
Q, R, S, U, V, W	Analysis of Realistic Scenes

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

Dataset

The dataset download links are available on Moodle. 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.
The features of the **4 additional scenes** (used for qualitative analysis) are provided in separate files, which are also available on Moodle (**Task 2: Data Exploration > Step 2: Download the dataset**). The structure of the file is equal to that of the training features. The raw waveforms (for listening) are also included in the zip file.

Grading

The written reports for this task and its subtasks are evaluated according to the following criteria

- **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 on time. Any delay will result in reduced grades. Specifically, submitting on May 23 will deduct 1/3 of the points, submitting on May 24 will deduct 2/3 of the points, and submissions on May 25 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 3 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. 7 pages where at most 4 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 2: Submit your team's Task 3 reports*) by **May 22nd**.
- You will get a maximum number of 37 points for your written report and 3 points for the slide set.