Project Assignment Winter Call, A.Y. 2022/2023

Alessandro Ficca
Politecnico di Torino
Student id: s319084
s319084@studenti.polito.it

***Abstract*—This report introduces an initial approach to speech recognition and classification. This classification task was based on audio files containing commands pronounced by different speakers. We have analyzed audio by extracting the mel spectogram and using a random forest and SVM classifiers.**

I. Problem overview

The dataset available is a collection of audio recordings of commands from different speakers. Each command is classified through an action and an object. Examples of commands are ”increase-volume”, ”decrease-volume” and ”increase-heat”.

The records contain the path of the audio file, the speaker Id and the speaker’s information like self-reported fluency level, first language spoken, the current language used for work/school, gender and age range. The class of the audio is obtained by the union of action and object features.

The dataset is composed by:

- A *development set*, containing 9854 records
- An *Evaluation set* of 1455 recordings.

We have used the development set to build a model to label each recording with an action and an object. The first analysis concerns the speaker’s information, we have studied the presence of correlations between the class assigned (action and object) and information available about the speaker, as we could have guessed there is no correlation between the features and the classes. Nevertheless, we noticed that some values are predominant in the dataset. For example, most of the speakers are characterized by a *”native”* self-reported fluency level, *”English (United States)”*as the first language spoken and the same for the current language used.

In our development set, there are 7 possible labels, but the classification task is unbalanced since the frequency of the classes is not well distributed. Figure 1 shows the frequency of the labels in the development set.

Then we analyzed the audio files. All recordings have been sampled at a frequency of 22050 Hz, which can be considered a reasonable frequency for speech recognition tasks. The duration of all audio differs, in figure 2 we can see that most recordings have a duration between 0.5 and 5 seconds, with the presence of some outliers which are up to 20 seconds long. The different lengths of recordings lead to a different number of samples for each audio since we have used the same sample frequency. So it is important to extract an equal number of features for each record.

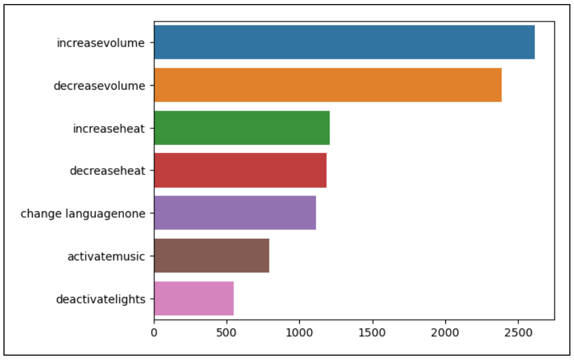

Fig. 1. frequency of classes in the development set

There are different ways to extract information from an audio signal. An audio file is a signal spreading over time, so it is possible to represent it as a time series in the time domain. The time domain permits clearly distinguishing how the signal changes over time. Another possible solution is representing a signal in the frequency domain, which shows how much of the signal lies within each given frequency band over a range of frequencies. The Fourier transform allows switching from the time domain to the frequency domain.

In our classification model, we have used the mel spectrogram. A spectrogram is a visual representation of the spectrum of frequencies of a signal over time, it is very useful for non-periodic signals, where the spectrum of frequencies is not constant [1]. The mel scale is a scale of pitches of sound, proposed in 1937, such that equal distances in pitch sounded equally distant to the listener. The mel scale exploits a logarithm scale and this is reasonable since humans do not perceive frequencies on a linear scale. [2] So the Mel Spectrogram is given by the following steps:

- Transformation from the time domain to the frequency domain using the fast Fourier transform
- Conversion of the frequency to a log scale and the amplitude to a decibel to form the spectrogram
- Frequency are mapped into the mel scale to form the mel spectrogram

Figure 4 is an example of a mel spectrogram extracted from an audio signal.

II. Proposed approach

A. Preprocessing

The preprocessing phase was very crucial in this task. It was composed by the following steps:

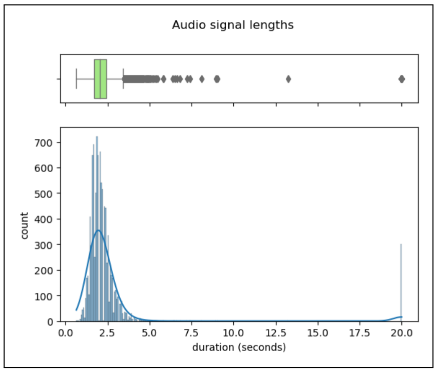

Fig. 2. Distribution of duration of recordings

- *Sampling of the data:* as we have said the problem was unbalanced (some classes were predominant), so we have resampled records from the classes with fewer records. With this approach, each class was represented by the same number of data.
- *Silence removal:* we have observed that some recordings contain periods of silence, so we have removed samples with amplitude smaller than 20 dB from the signals. Figure [3] shows how the length of the recordings was reduced.
- *Feature extraction:* We computed the mel spectrogram for each audio. To extract an equal number of features regardless of the duration, we divided the spectrogram into a fixed number of blocks. The number of blocks depends on the number of split along rows (N) and columns (M). For each block, we extracted the mean and the standard deviation of the values inside the blocks.

In this way, we obtained several features equal to $2 * M * N$ for each recording. At this point, we decided to add the categorical features encoded. We replace the possible values with integer numbers.

B. Model selection

The classifiers chosen for the task were the Random Forest and the SVM.
The Random Forest permits achieving an adequate accuracy score without excessive computational effort. With a balance in the distribution of the classes, the performance improves. It does not require normalization of the data. Whereas the performance of the SVM might increase with this process. Therefore we used a Z-score normalization on the features before training the SVM model.

C. Hyperparameters tuning

The first hyperparameters tune concerns the number of blocks extracted by the mel spectrogram. To simplify the computation, we choose the same number of splits along rows and columns, so $N = M = n$. Here are reported the different values of n tried:

- $n = 6 \rightarrow 30$, step 4

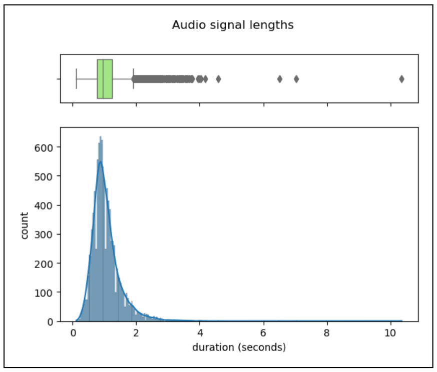

Fig. 3. Distribution of duration of recordings after removal of silence

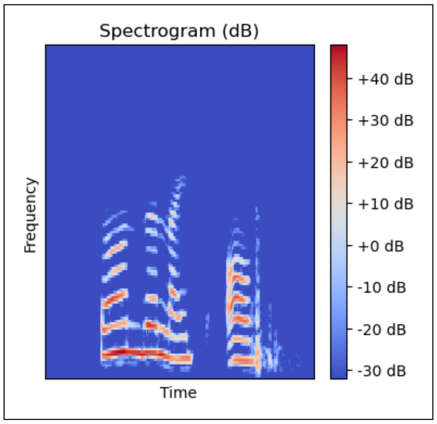

Fig. 4. Example of mel-spectrogram of an audio signal

The values of accuracy are obtained with a random forest with a default configuration. They are shown in figure 5.
In the second step, we looked for a better configuration of the random forest and the SVM. We have run a grid search on both algorithms with the following hyperparameters:

- Random forest
  - *max depth* = { None,10,50,100 }
  - *number of estimators* = { 100, 200 }
  - *criterion* = { gini, entropy }
- SVM
  - *C* = { 10,20,50,100 }
  - *kernel* = rbf

The grid search has been run with an 80/20 split on the development set to obtain the train/test sets. In the train/test split, we have distributed the data according to the feature ”First language spoken”, since it is predominant in the dataset.

III. Results

Good results were obtained with $n = 10$ and $n = 14$, with a slight difference. We choose $n = 10$ since it offers a lower computational cost. It can be considered a reasonable value because it does not increase excessively the number of values extracted. With the hyperparameters tuning we have found the parameters that perform best: {*criterion* = *entropy*, *max depth* = *50*, *n estimators* = *200*} and {*C* = *20*,*kernel* = *rbf*},

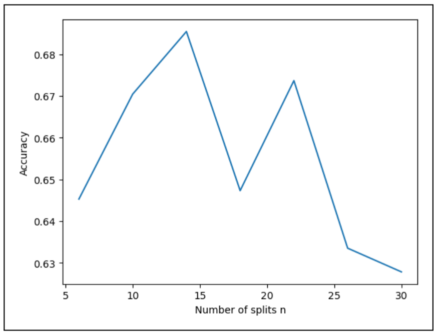

Fig. 5. Accuracy obtained with different values of n

respectively for the random forest and the SVM. However, In some cases, the differences in accuracy were slight.
These configurations have achieved an accuracy score of 0.90 for the random forest and 0.93 for the SVM in the development set. Then we trained the models on all the public datasets and used the models for the classification task on the evaluation datasets. On the evaluation set, the performances are quite different: The random forest reaches a score of 0.8, whereas the SVM has a score of 0.87. Overall, the SVM classifier leads to better results.

IV. Discussion

The results achieved can be considered quite good. Others approaches can be considered. For the feature extraction, an alternative way can be the Mel Frequency Cepstral Coefficients (MFCC). However, in our problem, this kind of approach has not led to a significant improvement.

References

[1] “Spectrogram — Wikipedia, the free encyclopedia,” 2010.
[2] “Mel scale — Wikipedia, the free encyclopedia,” 2010.