PROJECT 4: Acoustic Scene Classification

Elena Molinari

Melania Moras

Paolo Stella

Enrico Torres

1. Introduction
   1. Descripion of the task

The assigned task is Acoustic scene classification (ASC), whose purpose is to assign a label to an audio stream recorded in a certain environment by analyzing the audio signal. The data consists of 30-second audio files (WAV, stereo, 44.1 kHz, 16-bit), recorded using binaural headphones in locations around London at various times in 2012. In total, there are 200 audios equally distributed between ten different environments, which are: office, restaurant, bus, quiet street, busy street, tube, tube station, park, supermarket, open air market.

* 1. Characteristics of ASC

The dataset for this task exhibits a greater diversity and complexity with respect to speech or music signals. First, the length of the signals (30 seconds) suggests that in order to correctly identify the acoustic environment it is required to analyze long-term characteristics, in order to identify temporal patterns in the signals. For example, we can find many sound events over time or at the same time, and it is important not only to correctly identify them but also to analyze their characteristics. For example, we can find speech in both open air market and office, but it will be much quieter in the latter environment. Urban sounds are of course present in almost every signal, since they are all recorded in a city. We can assume that we will need a large number of features for the classification in order to capture the differences between each acoustic scene.

* 1. Literature review

ASC literature dates back to the 1970s, with the first work on the topic being soundscape (Schafer, 1977), pioneering work in which researchers pay attention to acoustic environments. Later, multiple ASC related concepts were proposed, and the automatic classification of environmental sounds received some attention. In the second phase, the publication of “computational auditory scene analysis” (Wang & Brown, 2006) and the study of comprehensively evaluating computers’ and humans’ perception of audio context (Eronen et al., 2006) started the early works of ASC. In this phase, ASC algorithms applied features from speech recognition areas and simple machine learning methods, as well as several ASC datasets were released. The third phase started from the DCASE2016 challenge (Eghbal-Zadeh et al., 2016), where deep learning methods were introduced into ASC and achieved high performance. In this phase, ASC received widespread attention, and the ASC publications rapidly increased. The performance of the research obtained some remarkable results on the LITIS dataset, with an accuracy of 98,9% in 2021.

Early methods were mainly based on MFCC features and simple machine learning methods, such as random forest, MLP, and SVM. Recently, features like LMBE, spectrogram, CQT, and log Mel spectrogram (or their variants) were adopted with CNN-based classifiers, such as ResNet, FCNN, and Xception. Feature learning combined with a simple classifier (e.g., MLP) also performs well in some datasets. Furthermore, most high-performance approaches employ data augmentation and late fusion methods, especially for processing complex datasets (e.g., D17 to D23).

Although the ASC accuracy on early datasets has been significantly improved to over 95 %, the accuracy on recent datasets is still low (less than 90 %), such as D20A, D21A, D22, and D23 datasets recorded by multiple devices. This is also due to the increase of task complexity (e.g., larger-scale data, multiple devices, limited model complexity, and shorter segments). Compared with the D20A task, D21A only added a low-complexity constraint, slightly decreasing accuracy.

* 1. Problems

Although the ASC field has made great progress, there are also some problems that need to be addressed. We analyze them from the perspectives of acoustic scenes, datasets, classification performance, and deployment in the real world

1. ***Unclear definition of acoustic scenes***. ASC requires a clear definition of acoustic scenes, as it needs predefined labels. However, it is hard to define a generalized boundary for acoustic scenes, since the acoustic signal is complex, multi-source, overlapping, and unstructured. Furthermore, the distribution of sound source components varies from application-to-application scenarios. For example, human activities and traffic are very intensive in urban environments, distinct from rural environments.

2. ***Lack of large-scale comprehensive datasets***. As a complex multi-classification task, a high-performance ASC system requires a large-scale dataset for model training. However, the size and diversity of the available data are currently limited. Large-scale comprehensive datasets are lacking for developing a specific ASC system, despite many existing ASC datasets.

3. ***Low classification performance***. Compared with other audio fields, the ASC’s performance is still deficient. There are mainly three reasons: (i) The high diversity among acoustic features of the same scenes and the similarity among acoustic features of different scenes make it difficult to extract efficient features for ASC, (ii) Audio signals lack structural information, making modeling difficult compared with speech signals, and (iii) Mismatched data distribution between source and target data, such as multiple devices, might cause low performance.

4. ***Computational efficiency and practicability***. At present, ASC models with high performance are relatively complex. There are problems with computational efficiency optimization and algorithm adaptability when deploying ASC algorithms in the real world. For example, in the reality of embedded industrial applications, an ASC system might be deployed in conditions of limited computational power and memory capacity (e.g., IoT embedded devices), where the high performance of the ASC system is also required. Therefore, the computational efficiency of the ASC system needs to be considered to judge whether it is practical for this application.

1. Methodology
   1. Initial steps

We worked with both the datasets available on the website, since the second one was initially private and then made public for testing. We merged the two folders into a single dataset, deciding to keep 50 samples for testing and perform data augmentation on the other 150, making them our training set.

First, we plotted the signals in order to gain familiarity with them, and we also plotted their Mel Spectrogram, in order to find differences between them. As expected, loud and quiet scenes have visible differences, while loud environments such as tube and bus do not have visible peculiarities. This confirms the hypothesis that using a small number of features will not result in accurate performances.

* 1. Preprocessing

The first preprocessing step is data augmentation. The data augmentation techniques that we employed are: time stretch, time compression, pitch shifting (up or down). In order to have the same number of samples per signals also for the augmented files, we had two options for the time stretched/compressed files, which were: zero padding and truncation or resampling. We opted for the latter, and we performed the four augmentation for each of the train files. We discovered soon that the stretched and compressed data compromised the results of the classification, so we eventually decided to discard them, keeping only the original files along with their pitch shifted versions.

Then we applied a scaler onto the signals, in order to normalize them.

* 1. Feature extraction

As a first approach, we created a feature vector containing the mean and standard deviation of the following quantities:

1. Zero-crossing rate
2. Mel spectrogram
3. Derivative of the Mel spectrogram
4. Second derivative of the Mel spectrogram
5. Spectral flux
6. MFCC
   1. Model

We opted for an SVM model, since previous reasearches in the literature seemed to obtain good results with it. The best kernel resulted to be ‘linear’, and the best value of C is 100.