SGN-16006, REPORT FOR EXERCISE ”COMPUTATIONAL AUDITORY SCENE RECOGNITION: CLASSIFICATION OF ENVIRONMENTAL SOUNDS”

Satu Haikonen, Juho Laukkanen

209491, 218886
satu.haikonen@helsinki.fi, juho.laukkanen@student.tut.fi

ABSTRACT

The aim of this MATLAB laboratory work was to build a simple classification system for the recognition of auditory environments. At first framewise processing of audio signals and feature extraction was studied in Assignment 1. Then the actual system for audio environment recognition was implemented in Assignment 2. Framewisely averaged sub-band energy ratio was used as the classification feature. For recognition, k-nearest neighbor (k-NN) classification was used.

1. INTRODUCTION

Classification problem refers to the problem of identifying in which class a set of new observation belongs to by comparing previous observations with the new observation [1]. In this laboratory report, we provide understanding into classification of environmental sounds by framewise processing of audio signals and extracting features that are designed to be informative and non-redundant data by reducing the dimensionality of the audio signals, thereby reducing the computation time at a given task.

Typically in audio signal processing the input data that is large and suspected to carry redundant data. Framewise processing introduces a method of dividing an audio signal into short, possibly overlapping segments called frames, lengths which are typically of tens of milliseconds [2, 3]. For each of these frames we can process and extract information from them individually.

For classification tasks extracting important information that describes the large data is called feature extraction [1]. Feature selection introduces methods or metrics that best differentiates audio signals from each other, be it energy-ratio, zero-crossing rate, fundamental frequency or spectral peaks[2]. By selecting which feature to use is at the root of the classification accuracy for any given classification task. Selecting features that are closely related between each signal the classification result might not end up being accurate at all. For the betterment of the classification task, audio signal analysis provides methods that are useful for understanding which features best differentiate each signal.

Selecting which classifier to use in practice is usually a matter of testing multiple different classifiers and seeing which of them provide the best result. Selection of a classifier goes hand in hand with the selection of which features to extract. Features that have proven to classify well with a certain classifier might not work as well on other classifiers.

This exercise was given in two separate parts, as an introduction as to how to perform framewise processing and extraction for one signal. Second task was to generalize processes learned from the first task and to perform them on multiple audio signals as well as to implement a classifier over the sample data after dividing it into training and testing data.

2. THEORY/BACKGROUND

For framewise feature extraction, we divide the signal first into smaller frames and apply a windowing function over all the frames. Window function is described as a mathematical function that is zero-valued outside a chosen interval [4]. Window function slides over the audio signal and segments it into adjacent not usually more than 50% overlapping samples. Some of the most commonly used window functions are the rectangular-, Hamming- and Hanning-window functions.

To extract features, we process each window function frame individually and calculate a pre-defined feature. The assigment introduces sub-band energy ratio which describes the relative energy at certain frequency bands. As a result of the Nyquist theorem the sub-band division is performed on half of the sampling frequency. For each sub-band an energy ratio is calculated. If we denote the discrete Fourier transform (DFT) of a signal frame as S following the equation 1:

$$x(i) = \frac{\sum_{l=b_i}^{e_i} |S(l)|^2}{\sum_{l=0}^{L/2} |S(l)|^2} \tag{1}$$

where L is the number of frequency bins defined by the calculation of DFT and b_i and e_i denote the first and last DFT bins that belong to the ith frequency band [3].

The k-NN or k-Nearest Neighbors classifier performs a class vote for a new data point and defines its class on the basis of its closest k nearest neighbor. Different distance metrics are used, but the commonly used for continuous data is Euclidean distance [1, 3]. Classification accuracy can be improved if the distance metric is optimally selected by using for example Neighbourhood components analysis.

3. METHODS

In the first assignment, at first the example signal was read from a wav file. The audio clip was divided into frames by using Hanning window with frame length of 30 ms and overlap of 15 ms of adjacent frames. The sub-band energy ratio for the 50th Hanning-windowed frame was calculated with the help of discrete Fourier transform (MATLAB function fft). The number of bins was set to 1024.The sub-band energy ratios were calculated for frequency bands [0-0.5, 0.5-1, 1-2, 2-4]kHz.

In the second assignment the data consisted of 8 kHz MP3-recordings of 5 minutes each in separate wav files from 12 different audio environments including a building site, bus, office, shopping mall, etc. The files were read in MATLAB and segmented into one-second clips without any overlap between the audio clips. The short clips were then divided among training and test data sets so that approximately 80 % of the data from each acoustic environment was included in the training data and 20 % in the test data.

Each training data clip was processed framewise and the sub-band energy ratio values were extracted for each frame using Hanning windowing and fft (similarly as in the first assignment). The average of each subband energy ratio value over the frames of the audio clip was computed resulting in one average feature vector of size 4x1 for each audio clip. Both test data and train data were processed similarly.

In the classification phase of the second assignment, a NN classifier predicting the environment class index of a test audio clip based on its first nearest neighbor among the training samples was built. The idea was generalized to a k-NN classifier and the prediction accuracies of each class were computed for both classifiers.

4. RESULTS AND DISCUSSION

In the first assignment the sampling frequency of the one-second audio signal was 8 kHz. The length of one frame in samples was 240 samples. The plot of the waveform of the 50th Hanning-windowed frame is shown in Fig. 1. The amplitude spectrum of the same signal frame is presented in Fig. 2. The subband energy ratios of the frame for the frequency bands: [0-0.5, 0.5-1, 1-2, 2-4]kHz were 0.3433, 0.3501, 0.2229 and 0.0837 respectively. The indices of the DFT bins belonging to the frequency band 12 kHz were 129-256.

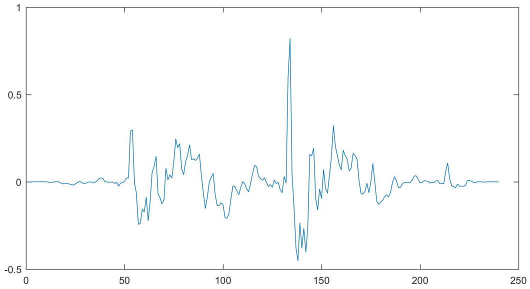

Fig. 1. The waveform of the 50th Hanning-windowed frame of an example one-second audio signal.

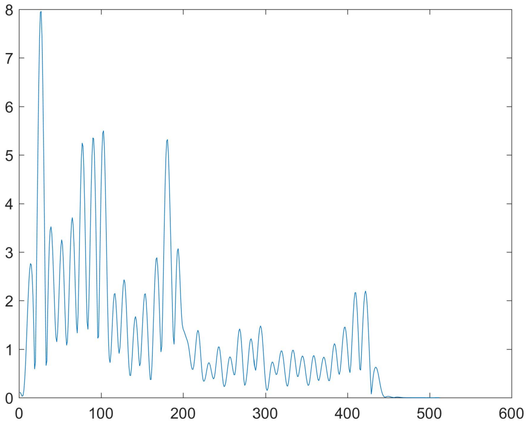

Fig. 2. The amplitude spectrum of the 50th Hanning windowed frame.

The results of the classification are shown as confusion matrices in Table 1 for 1-Nearest Neighbor classifier and in Table 2 for 5-Nearest Neighbor classifier. The classification accuracy for each class with both 1-NN and 5-NN classifier are shown in Table 3.

From the confusion matrices can be seen that the classification accuracies are relatively good for both classifiers. The class-wisely presented accuracies also suggest the same. The different environments starting from number one were buildingsite, bus, car, car on a highway, laundrette, office, presentation, shopping centre, people on the street, traffic on the street, supermarket and train.

As can be seen from the accuracy numbers on Tables 1, 2 and 3, the office environment (class number 6), was the ”easiest” to classify, as both classifiers had 100 % accuracy in relation to it. On the other hand, the accuracies were lowest for shopping centre (class 8) and supermarket (class 11), suggesting that they were the most difficult to classify. The

Table 1. Confusion matrix of 1-Nearest Neighbor

Class	1	2	3	4	5	6	7	8	9	10	11	12
1	57	0	0	0	0	0	0	0	0	1	0	0
2	0	38	10	11	1	0	0	1	3	1	0	0
3	0	4	41	1	0	0	4	0	0	0	0	0
4	0	9	3	42	0	0	6	0	0	0	0	0
5	0	3	0	1	44	0	3	2	5	0	3	3
6	0	0	0	0	0	56	0	0	0	0	0	0
7	1	7	6	4	4	0	44	0	0	0	0	0
8	0	0	0	0	1	0	0	24	7	9	10	13
9	0	12	0	1	3	0	0	2	26	3	8	0
10	0	0	0	0	0	0	0	4	1	52	4	0
11	0	1	0	0	4	0	0	10	2	6	26	7
12	0	0	0	0	0	0	0	8	5	1	3	48

Table 2. Confusion matrix of 5-Nearest Neighbor

Classes	1	2	3	4	5	6	7	8	9	10	11	12
1	57	0	0	0	0	0	0	0	0	1	0	0
2	0	41	11	10	0	0	1	0	2	0	0	0
3	0	4	41	1	0	0	4	0	0	0	0	0
4	0	6	3	45	1	0	4	0	0	0	0	0
5	0	4	0	0	47	0	1	1	6	0	3	2
6	0	0	0	0	0	56	0	0	0	0	0	0
7	1	6	5	10	4	0	40	0	0	0	0	0
8	0	0	0	0	1	0	0	29	6	6	4	18
9	0	12	0	1	1	0	1	5	28	1	6	0
10	0	0	0	0	0	0	0	7	0	51	3	0
11	0	3	0	0	1	0	0	15	5	5	25	2
12	0	0	0	0	0	0	0	9	1	0	2	53

Table 3. Classification accuracies for each class

Class	Accuracy 1-NN	Accuracy 5-NN
1	0.9828	0.9828
2	0.5846	0.6308
3	0.8200	0.8200
4	0.7000	0.7500
5	0.6875	0.7344
6	1	1
7	0.6667	0.6061
8	0.3750	0.4531
9	0.4727	0.5091
10	0.8525	0.8361
11	0.4643	0.4464
12	0.7385	0.8154

accuracies for both shopping centre and supermarket were below 0.5, which is theoretically the accuracy of guessing.

5. CONCLUSION

In this work, a simple classification system for the recognition of auditory environments was implemented. In the training phase, the implemented program read in one-second audio clips and their correct class labels, extracted framewise energy features, averaged them over the audio clip, and store the values and labels in MATLAB memory. In the test phase, the program read in a test clip, carried out the feature extraction and averaging over frames and to predict the class label based on the k nearest neighbors from the training data.

The accuracy of the prediction was also evaluated. The accuracies of k=5 classifier appeared slightly higher than the accuracies of k=1 classifier. We evaluated that we had good results for accuracy, but in realistic applications the code might have to be optimized to perform fast enough.

We definitely think that the skills learned during with this laboratory work will be useful in the future. The idea of this laboratory work was very logical, sensible and interesting. However, we had a few problems understanding the work instructions, which were at some point unambiguous. Especially the concept of computing the average feature vectors (3.2, task 2) was from our point of view unambiguously written, though a simple task in the end. The amount of time used for this work including this report was approximately 30 hours per person.

REFERENCES

[1] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.

[2] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition." in *Proceedings of ICASSP*, 2002.

[3] "Audio processing work instructions, spring semester 2016," *Bachelor's Laboratory Course in Signal Processing SGN-16006*.

[4] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Addison-Wesley, 2005.