Introduction to Audio Processing – Project Assignment

Environmental Audio Analysis

Author

Name: Sijan Pandey
Student number: H293831
Email address: sijan.pandey@tuni.fi

1. Introduction

In this work, audio clips taken from everyday environments involving different scenes and situations were analyzed. This work explores the average similarity of audio clips labelled with same labels. To achieve this, acoustic features of audio clips with different labels were compared.

2. Data annotation process

2.1 Annotation process

Audio clips from everyday environments involving different scenes and situations were annotated. There were altogether nine labels available. Each audio clip could be labelled with multiple labels. If the clip didn't match with the any labels given, it was possible for the annotator to provide self-created labels. However, these self-created labels weren't considered during audio analysis. Annotator had to provide short description of the audios. These descriptions were not used during the analysis.

The annotator tool was quite easy to use and there was no lengthy log-in process to access the tool. It was good that annotation could be resumed from where you left off. However, on a negative note, recognizing every sound in the clips was difficult with normal headphones even in full volumes. Subtle sounds were difficult to recognize. Some labels might have been missed due to that reason. In addition, due to human nature, sometimes environmental random noise might have been labelled mistakenly assuming it to be footsteps or traffic noise.

Further improvement could be done on the audio collection so that even subtle sounds are captured well.

2.2 Dataset statistics

Table 1 lists some of the statistics of the dataset provided for analyzing.

Table 1: Dataset statistics

Number of audio samples	131
Length of each audio sample	10 seconds
Number of classes present	9
Average number of labels per file	1.748
Most frequent label	adults_talking (84 audio samples)
Least frequent label	siren, announcement_speech (3 audio samples)
Number of unlabelled samples	4

Table 1 shows that most of the labels are multi-labelled. There is imbalance of classes in the dataset. Hence, less frequent labels will provide unreliable results. There few unlabeled samples so it wouldn't skew the results a lot.

3. Audio analysis

3.1 Implementation

MFCC of the audio samples were calculated with 40 mel filters, hop length of 512 and fft size of 1024. This setting accounts for the length of each frame to be approximately 20 ms and the

MFCC were calculated with 50% overlap. Shorter frames are usually suitable for normal speech like audios. Hence, approximately 20 ms frames were selected. Different number of filters, both higher and lower, were experimented. However, the resulting similarity matrix value didn't seem to change much. Hence, 40 mel filters (default as in the example in project instructions) were used. Mean and standard deviation of temporal axis of MFCC matrix were taken for each filter and these were concatenated to form a vector of dimension 80x1. These were the feature vectors of the audio samples. For the analysis, feature vectors of the audio samples were compared using cosine similarity. A similarity matrix was generated showing similarity metrics relationship between all the samples.

3.2 Results and discussion

Distance is a quantity used to quantify how far two vectors are in the hyperplane. Similarity signifies how close two vectors are. Distance has value 0 for two same vectors and higher values for different vectors. Similarity (cosine similarity) has value 0 for two most dissimilar vectors and 1 for two same vectors. Figure 1 shows the plot of similarity matrix.

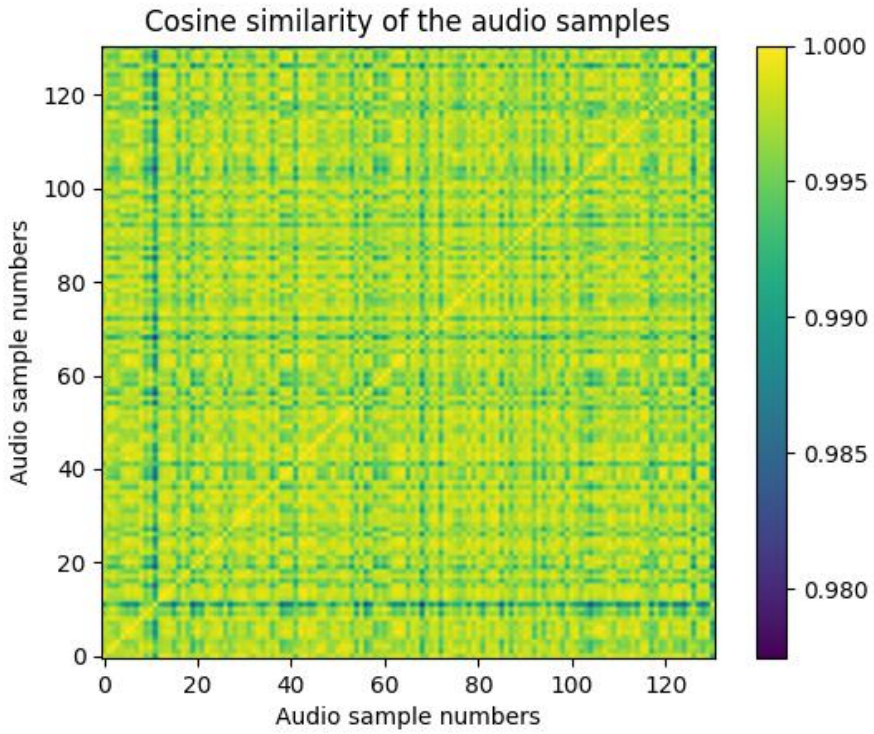

Figure 1: Cosine similarity matrix

It was also observed that the samples that are from same location aren't guaranteed to be most similar acoustically. Similarity is observed to be based on their acoustic contents. For example, first airport sample isn't similar to second airport because they contain audio contents of different classes. It was observed in other samples also.

The similarity matrix shows that all the audio clips are almost similar to each other except for few classes. It was observed that average of the similarity matrix is 0.9968276 which is almost equal to 1. The reason behind it can be that all the audio clips are from airport, park or public square. All of them are public places with people and other common animals which can share

almost similar audio contents. For example, even if there is a dog barking in a park, at the same time there can also be birds and children and adults over there. However, if these actors are in background and their sounds are subtle it can be easily missed during annotation process so these might look completely different samples based on labels but can have similar content.

Table 2: Average similarity for classes

Class	Average similarity for class
adults_talking	0.9967461
dog_barking	0.99853367
footsteps	0.9973734
traffic_noise	0.9971948
birds_singing	0.99799055
siren	0.9985394
music	0.99804914
children_voices	0.9956979
announcement_speech	0.9964023

It can be observed that announcement, adults_talking and children labels have similarity values lower than that of the entire matrix. This signifies that audio features of those classes are quite different in different samples and the feature vector doesn't represent them well. Number of children (13) and announcement (3) labelled samples are few, so this result isn't conclusive for them. However, this is good evidence for adults_talking samples. In other labels, acoustic patterns are more similar regardless of having multiple labels which means that the audio features are more consistent based on the labels even if the samples are multilabelled. This also means audio features of these labels are prominent in the feature vector. It is also interesting to mention that single labelled footstep samples have high similarity with other samples labelled as footsteps and even with audio samples that aren't labelled as footsteps but other labels like siren, traffic noise. In fact, footsteps labelled samples are highly similar to almost all other labelled samples except for non labelled samples and samples that are single labelled as only adults_talking or children_voices. However, they are found to be similar with adults_talking and children_voices samples that are multi-labelled with other labels. One reason behind it could be that the beats from other labelled samples could have similar features to that of footsteps samples.

4. Conclusions

Environmental audio samples were annotated with labels. Annotated samples were analyzed using feature vectors produced from MFCC of samples. The samples were found to be a lot similar to each other despite of having different labels based on this method of comparison.