Audio Processing Assignment Report - Classification of Audio
Vasav Juyal, Miro Malekshahi

Introduction

This report deals with describing the process involved in binary classification of sounds by classifying audio files into predefined categories using a series of systematic steps. This involves collecting and processing audio data, extracting relevant features, and implementing a machine learning model to perform the classification. The model is used to predict from data it has never seen before, then the results from these predictions are evaluated to see how accurate the model is at classifying the audio. For this task, it was assumed that the audio samples used for training were accurately labeled and did not contain too much noise.

Data description

The data is audio collected from vehicles, mostly using the recording feature available on phones. The data was collected in Tampere, either by recording the vehicles as they passed or while they were idle. This approach ensured a diverse set of audio samples, typically around five seconds long, reflecting various vehicle states and conditions. The dataset is designed for binary classification, meaning it contains two distinct classes. The classes chosen here are car and tram. Our group collected around 25 samples each, and the rest of them were collected from freesound.org, an open source website that provides free access to sounds collected by others. Using this, we gathered 850 samples for the car class and 809 samples for the tram class.

There were many audio formats present in the collected samples, so they were all converted to wav files, to ensure consistency while handling them. Due to the nature of these samples, they were also normalized to ensure uniformity across the dataset. Normalization adjusts the amplitude of the audio files to a standard level, which helps in reducing variability caused by different recording conditions and devices. This step is crucial for improving the accuracy and reliability of the subsequent feature extraction and classification processes.

Feature extraction

The features serve an important role in the task. Due to the difficulty presented in trying to analyze raw audio data, extracting features that provide information about the audio is a more feasible way to provide a basis to the model for classification. Two time-domain and two frequency domain features are chosen here. Each feature provides unique insights into the audio data, and their selection is justified based on their relevance and effectiveness in capturing important characteristics of the sound.

The time-domain features used are Root Mean Square Energy (RMS) and Zero-Crossing Rate (ZCR). RMS energy measures the overall magnitude of a signal corresponding to its energy. For audio signals, this generally equates to how loud the signal is. It is particularly useful for distinguishing between different types of sounds based on their loudness [1]. This might be particularly useful for this task, as the tram might have a higher RMS energy due to its size. The ZCR feature gives us a count of how many times the wave crosses the x-axis. It is an important feature to take into account because it gives us a representation of the smoothness of the wave by calculating the number of times it changes from positive to negative and vice versa [2]. The class which has a more continuous sound will tend to have a lower ZCR, so it can be useful in identifying differences.

The frequency-domain features used are Mel-Frequency Cepstral Coefficients (MFCCs) and Spectral Centroid. MFCCs represent the short-term power spectrum of an audio signal on a mel scale, approximating the human ear's response to different frequencies [3]. They capture the timbre of the sound, which helps distinguish the harmonic structure of the engines. Spectral Centroid indicates where the majority of a signal's energy is concentrated.“Spectral centroid can be used to understand the brightness of the sound wave. Intuitively a certain collection of higher frequencies would make a wave sound brighter than if it were replaced by the same collection of frequencies in the lower octave” [2].

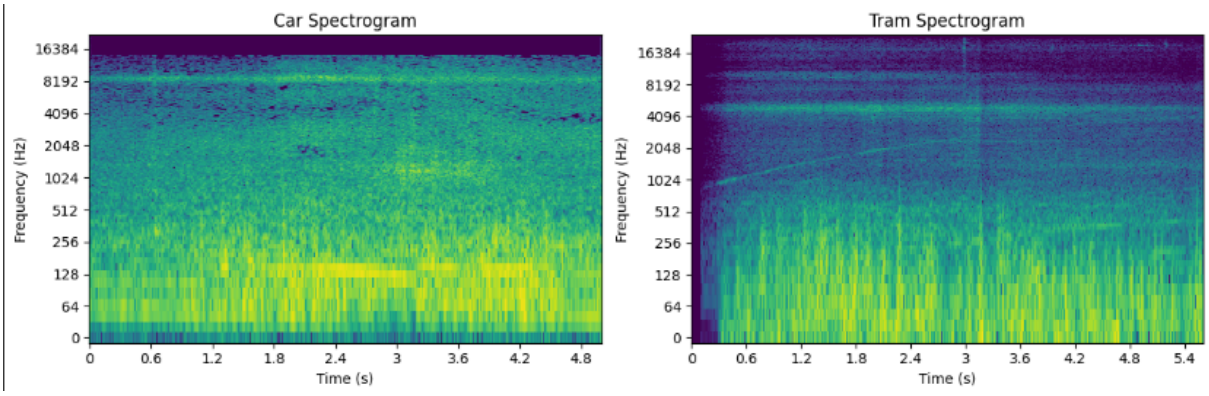

Fig 1: Spectrograms for a random sample of car and tram. Cars have higher energy at lower frequency and more consistent distribution and it's the opposite for trams.

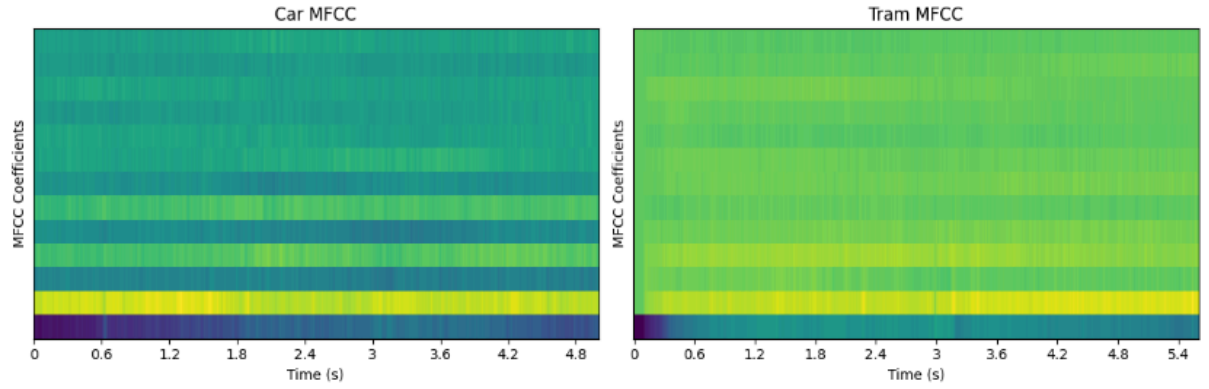

Fig 2: MFCCs for a random sample of cars and trams. Low frequencies are dominant for car and high frequencies are dominant for tram

MFCC scored highest on each of the classification metrics, which will be further discussed in the results section, so it was selected for use in the final model as the best feature for classification.

Model selection, data split

Logistic regression was the model chosen for the classification task here. Logistic regression is used for binary classification and predicts the probability of an instance belonging to a class. When there are only two classes, binomial regression is used, which predicts either 0 or 1 instead of values between 0 and 1 [4]. For training data, all the data collected from freesound was used. A total of 1505 samples were used in training, with 733 for tram and 772 for car. The test data consists of audio samples gathered by our group, 24 for tram and

29 for car. The validation data was collected from three separate users in freesound, who had recorded samples for both cars and trams. For validation data, the tram class had 52 samples and the car class had 49. First, a model is trained on each of the four selected features individually and used to predict on the validation data. The results are analyzed and the best feature is selected based on accuracy. Then a new logistic regression model is trained on the training and validation data, which is then used to predict the test data. None of the models used any hyperparameters, keeping the base version of the model and consistency for all data.

Results

The primary objective of this evaluation is to determine the effectiveness of different audio features in classifying vehicle sounds into two categories: car and tram. The features evaluated include Zero-Crossing Rate (ZCR), Root Mean Square Energy (RMS), Spectral Centroid, and Mel-Frequency Cepstral Coefficients (MFCCs). Each feature's performance is assessed using three metrics: accuracy, precision, and recall.

	Accuracy	Precision	Recall
ZCR	0.554455	0.600000	0.244898
RMS	0.613861	0.589286	0.673469
Spectral centroid	0.514851	0.500000	0.408163
MFCC	0.910891	0.857143	0.979592

Table 1: Performance Metrics for Different Features using validation data.

Accuracy represents the ratio of correct classifications to total classifications. Precision represents the ratio of correctly identified positives to all identified positives and recall is the ratio of correctly classified positives to all actual positives.

MFCC was the best feature to use for the classifier, as it scored highest on all metrics tested by a large margin. As seen in Table 1, not even the second place RMS comes close to the MFCC in any metric. The figures in table 1 are an average for car and tram samples.

	Precision	Recall	Accuracy
Car	1.0	1.0	1.0
Tram	1.0	1.0	1.0

Table 2: Performance of the MFCC model using test data.

As seen in table 2, the model performed perfectly on the test data when using MFCCs alone.

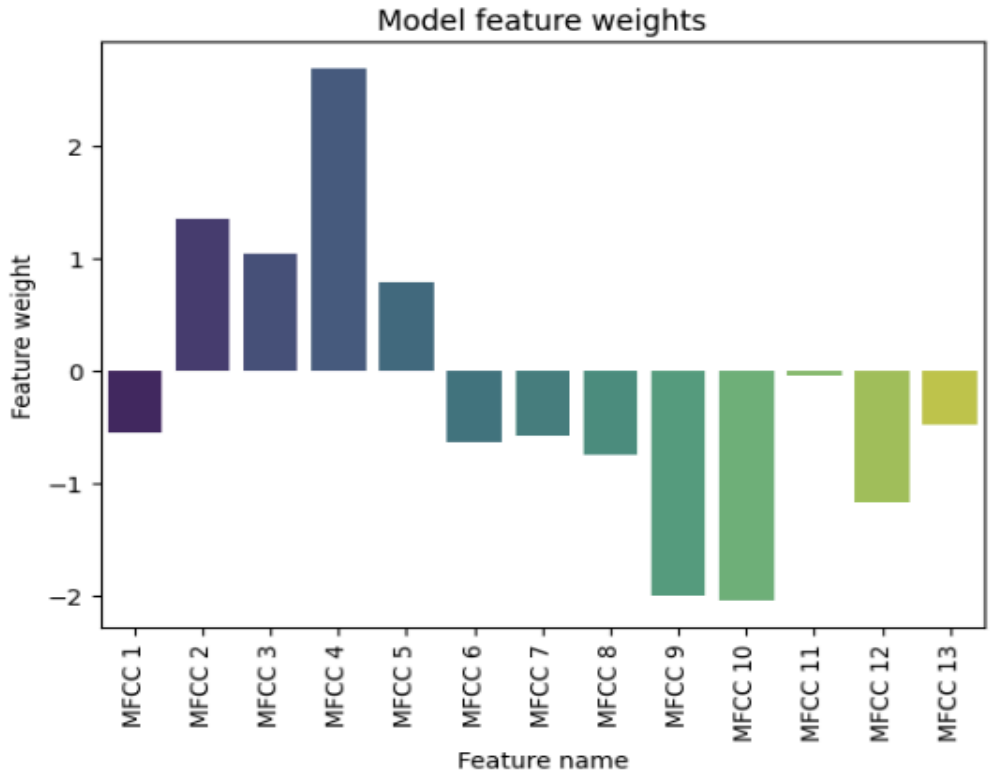

Fig 3: Weights of different MFCC bands used in the model for test data

Figure 3 illustrates the different weights assigned to the different MFCC bands in the model. MFCC bands 2 to 5 had the most influence on the model and 9, 10 and 12 had the least influence.

Conclusion

The execution of the project was straightforward, with no trouble encountered. When evaluating the model, MFCC was the best in every metric used for evaluation. This might be due to the fact that MFCC has several bands and thus more possible dimensions to be used for training than other features tested. As can be seen in Figure 3, not all bands are equally useful for classification. It could be that removing bands 9 and 10, which have the lowest weights, would not have a noticeable impact on the performance of the model. It could be that the other features outperform the lower weight bands in the model and would make the model better.

The perfect results in the test data may be due to the small amount of samples in the test data. During experimentation, it was found that ZCR performed better than RMS on the test data even though by most metrics RMS was better on the validation data. With less data to work with results can be skewed quite easily, and is likely the explanation for such unexpected accuracy.

Both group members participated in data collection, coding and writing the report. Each member collected the samples for a single sound category. In coding, one member worked on splitting the features and training the model, while the other worked on testing, evaluating and displaying the result. Both members worked on the report.

References

[1] Drishti. (2022, May 27). Comparison of the RMS energy and the amplitude envelope. Analytics Vidhya. Retrieved from Comparison of the RMS https://www.analyticsvidhya.com/blog/2022/05/comparison-of-the-rms-energy-and-the-amplitude-envelope/Energy and the Amplitude Envelope

[2] Dwivedi, D., Ganguly, A., & Haragopal, V. V. (2023). 6 - Contrast between simple and complex classification algorithms. In T. Goswami & G. R. Sinha (Eds.), *Statistical Modeling in Machine Learning* (pp. 93-110). Academic Press. https://doi.org/10.1016/B978-0-323-91776-6.00016-6

[3] Pengfei, L., & Dong, X. (2023). 12 - Prediction model of residual load-bearing capacity of composite laminates using deep learning. In P. Liu (Ed.), *Acoustic Emission Signal Analysis and Damage Mode Identification of Composite Wind Turbine Blades* (pp. 303-342). Elsevier. https://doi.org/10.1016/B978-0-323-88652-9.00002-9

[4] GeeksforGeeks. (2024, June 20). Understanding logistic regression. Retrieved from https://www.geeksforgeeks.org/understanding-logistic-regression/