**Classification of Urban Sounds using**

**Convolutional Neural Networks**

**1.1 Introduction**

Audio signal classification (ASC) consists of extracting relevant features from a sound, and of using these features to identify into which of a set of classes the sound is most likely to fit [1]. Audio classification has been of interest to the science and engineering community as it presents great potential for the implementation in a wide array of human computer interfacing systems. Such interfaces include but are not limited to speech to text, voice recognition, audio scene tagging and many more.

In this project I propose the sonic analysis of urban environments which has been enabled by multimedia sensor networks, as well as by large quantities of online multimedia content depicting urban scenes.

However, while there is a large body of research in related areas such as speech, music and bioacoustics, work on the analysis of urban acoustic environments is relatively scarce. Furthermore, when existent, it mostly focuses on the classification of auditory scene type, e.g. street, park, as opposed to the identification of sound sources in those scenes, e.g.car horn, engine idling, bird tweet.

There are primarily two major challenges with urban sound research namely

* Lack of labeled audio data. Previous work has focused on audio from carefully produced movies or television tracks from specific environments such as elevators or office spaces and on commercial or proprietary datasets . The large effort involved in manually annotating real-world data means datasets based on field recordings tend to be relatively small (e.g. the event detection dataset of the IEEE AASP Challenge consists of 24 recordings per each of 17 classes).
* Lack of common vocabulary when working on urban sounds. This means the classification of sounds into semantic groups may vary from study to study, making it hard to compare results so the objective of this notebook is to address the above two mentioned challenges.

**1.2 Project Description**

The purpose of the project is to classify sounds in the urban environment. A dataset containing labeled sound excerpts (<=4s) of urban sounds from 10 classes is used in this endeavor. The dataset contains 8732 sound excerpts (<=4s) of urban sounds from 10 classes, namely: Air Conditioner Car Horn Children Playing Dog Bark, Drilling, Engine Idling, Gun Shot, Jackhammer Siren, Street Music. As sounds vary widely in their frequency and time content the approach of using features of the sound that can be more effectively used to in classification is necessary. Supervised machine learning is used as a means of classification as a training and testing dataset are available to provide class labels and to test the efficiency of machine learning techniques. A Convolutional Neural Network is used to classify urban sounds after the relevant features of the sound signals are processed.

**1.3 Classification Overview**

There are many challenges related to the acoustics of classification. First, the acoustic characteristics of even a single class of sounds can be highly diverse. For example, in the case of class “baby crying,” the acoustics can vary enormously depending on the baby who is crying and the way in which they cry. Second, in realistic environments there can be many different types of sounds, some of whose acoustic characteristics may be very close to the target sounds. For example, the acoustics of a baby crying can be close to vocals in some overpowering noise source that is present in many environments. Thirdly, an audio signal captured by a microphone is affected by the channel coupling (impulse response) between the source and microphone, which may alter the signal sufficiently to prevent matching of models developed to recognize the sound. Finally, in realistic environments there are almost always multiple sources producing sound simultaneously. The captured audio is a superposition of all the sources present, which again distorts the signal captured. In several applications of sound classification, microphones that are used to capture audio are often significantly further away from target sources, which increases the effect of impulse responses from source to microphone as well as other sources in the environment. This situation is quite different from speech applications, where close-talk microphones are still predominantly used.

To address these problems a diversity of the training set present in the data is needed. Pre processing methods such as filtering based on the targeted sound characteristics may be needed. In this project the computational analysis dealing with target sounds are based on the *supervised machine learning* approach, where the system is trained using labeled examples of sounds from each of target sound type [2]. The supervised learning approach requires that there is a set of possible sounds event (e.g., airhorn, drill,car horn, dog barking etc) categories, *classes*, and that there is sufficient amount of labeled examples available to train the system. An overview of classification is shown in Fig 1.

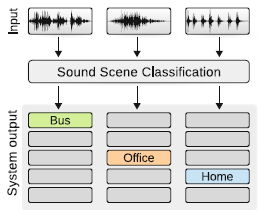


Figure 1. Sound Classification Schema

**1.4 Audio Processing**

Audio processing may be categorized into two stages. First the audio is *pre-processed* to remove unwanted characteristics pre-processing may include applying an inverse transformation of the environment the sound has been recorded in if it is known. For example an inverse room impulse response can be applied to the sound signal if the environment (room size and shape is known). Other pre processing can include filtering the sound signal if the underlying frequency range of the sound is known, this is done in order to filter out noise that may contaminate the signal. The second stage includes *feature extraction* which is applied in order to represent the signal in a more compact form.

**1.4.1 Audio Feature Extraction**

The purpose feature extraction is to transform the signal into a representation that improves the sound recognition performance of the sound classification system. The acoustic features provide a representation of the signal information for machine learning. The processing pipeline in feature extraction is similar for many types of acoustic features used in analysis and consists of *frame blocking*, *windowing*, *spectrum* calculation, and subsequent analysis, as illustrated in Fig. 2

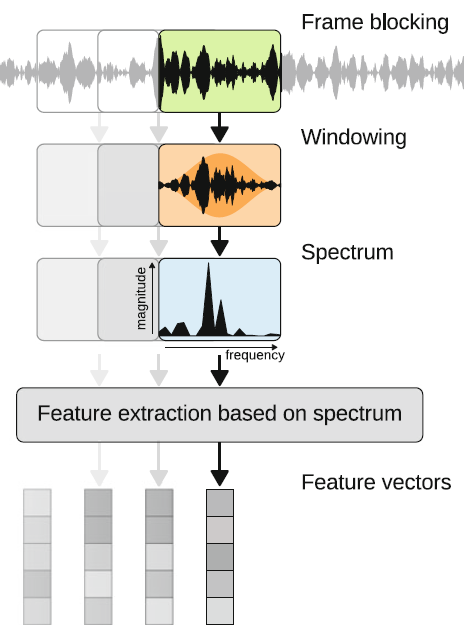


Figure 2. Feature Extraction Pipeline

Digital audio signals are discretized in terms of both amplitude and time when captured. For audio analysis, a significant amount of information is contained in relative distribution of energy in frequency, suggesting use of frequency domain features or time-frequency representations. The most common transformation used for audio signals is the discrete Fourier transform (DFT), which represents the signal with a superposition of sinusoidal base functions, each base being characterized by a magnitude and phase [3]. Examples of other transformation methods used for audio signals are constant-Q transform (CQT) [4] and discrete wavelet transform (DWT) [5].

The most common acoustic features used to represent spectral content of audio signals are mel-band energies and mel-frequency cepstral coefficients (MFCCs) [7]. The design of the mel-frequency cepstral coefficients are based on human auditory perception. This perception is based on the magnitudes of the frequency contents of sound.The perception of these magnitudes is non-linear, and the perception of frequencies is also non-linear. As a result a non-linear frequency scaling is implemented using filter banks which integrate the spectrum at non-linearly spaced frequency ranges, with narrow band-pass filters at low frequencies and with larger bandwidth at higher frequencies. This approach is used for classification of the 10 urban sounds in question.

**1.4.2 Acoustic Feature Extraction (Time Domain)**

Sound is the propagation of a material disturbance through a medium. A medium may be thought of as a volume of matter of a particular composition. Mediums are made up of particles which are assumed to be motionless relative to each other. The material properties of the medium such as its, density and bulk modulus help determine the velocity of the propagating sound wave. In short, a sound wave is the movement of a piece of information from one point to another in a medium [1]. The results of sound wave propagation are the compression and rarefaction. Compression is an increase in pressure within a section of a medium created when its composite particles are pushed together. Rarefaction is a decrease in pressure within a section of medium when particles are farther apart.

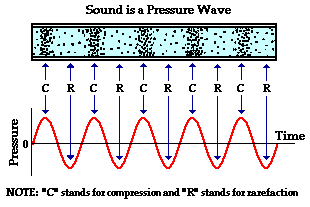
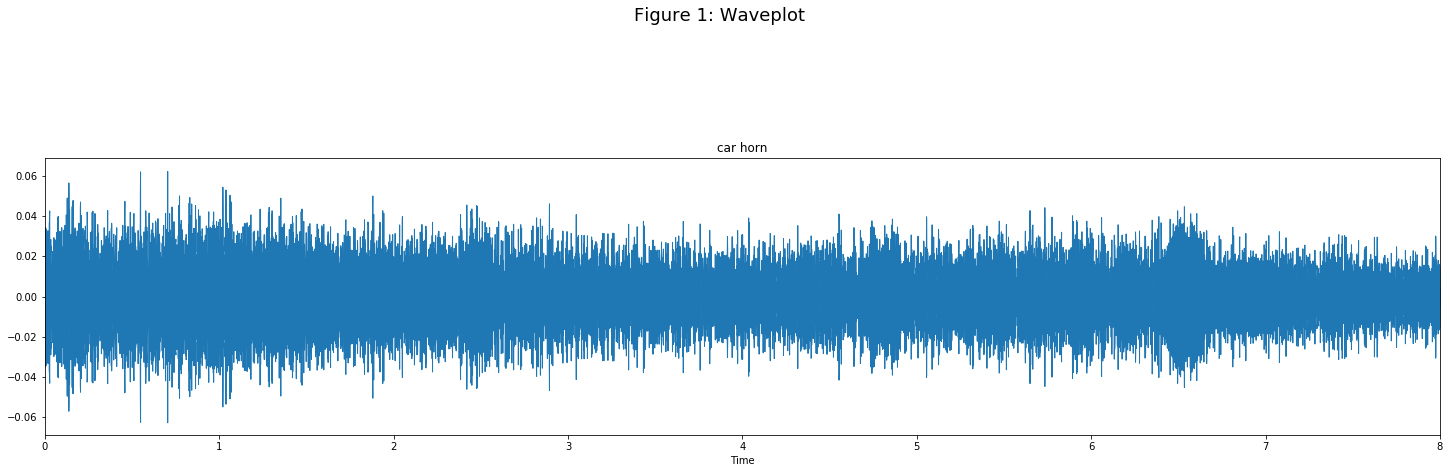
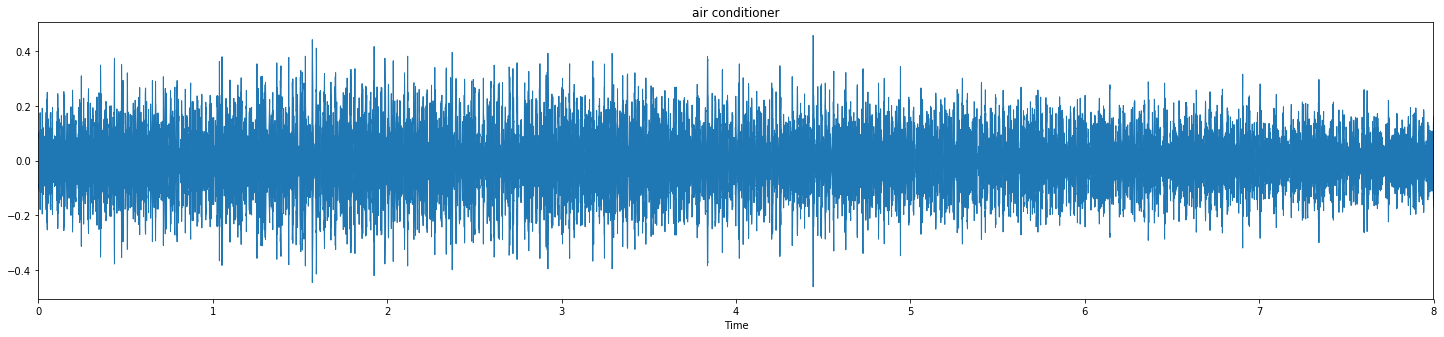
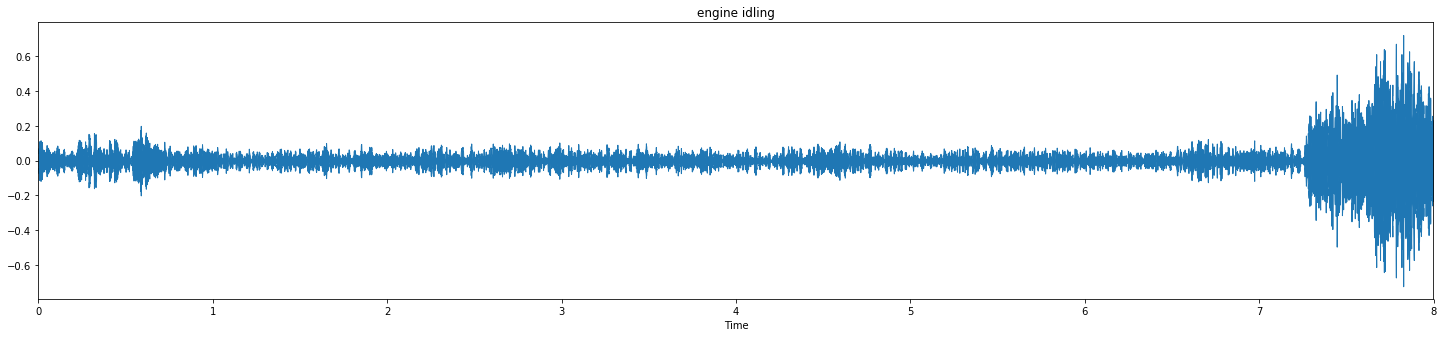
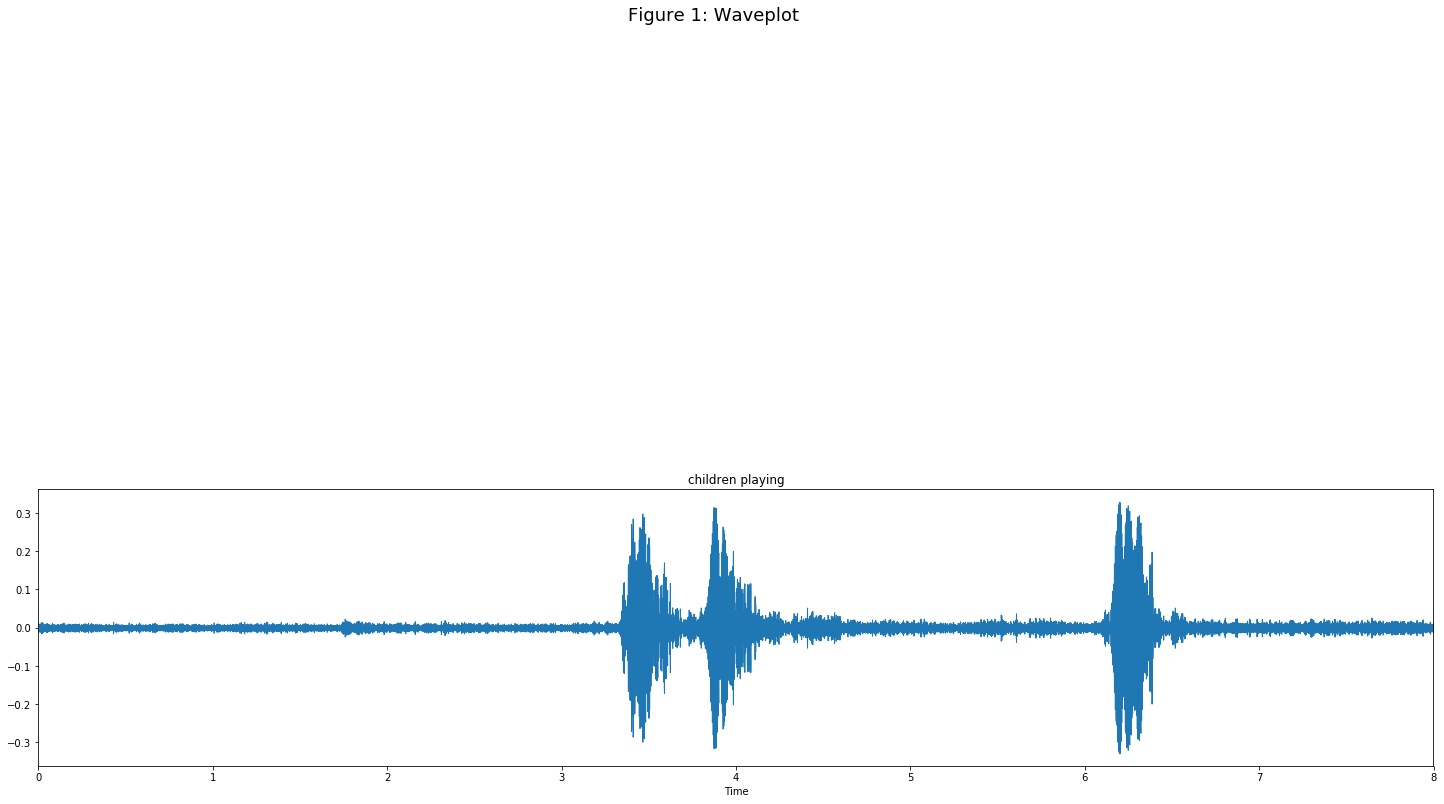


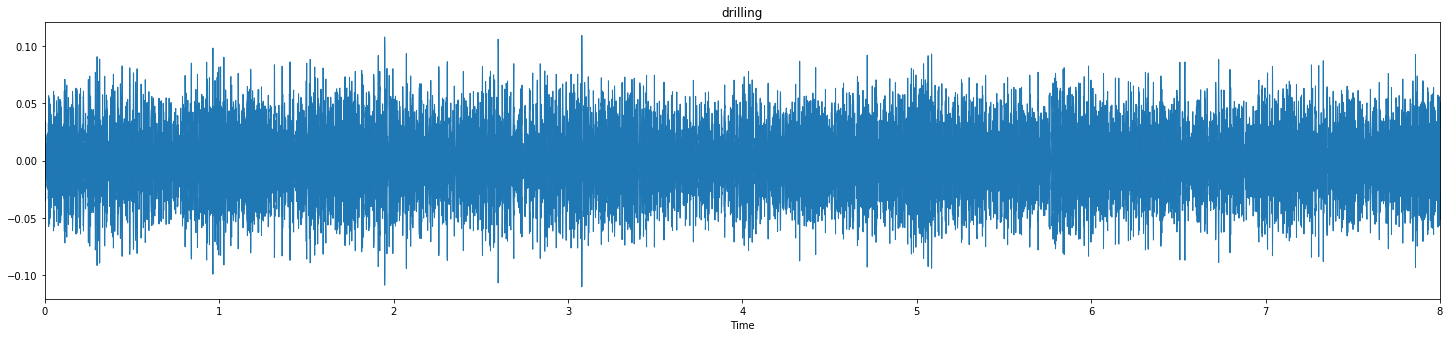
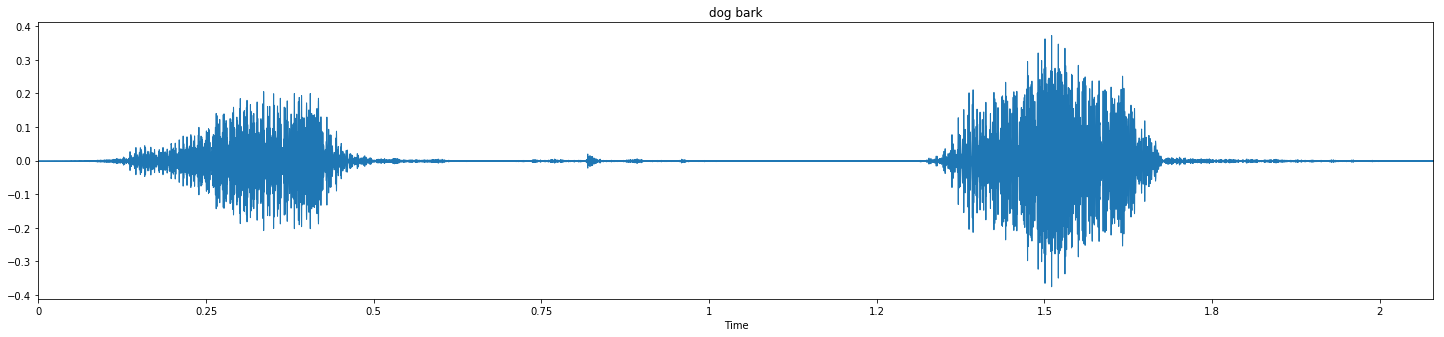
Figure 3. Compression and Rarefaction

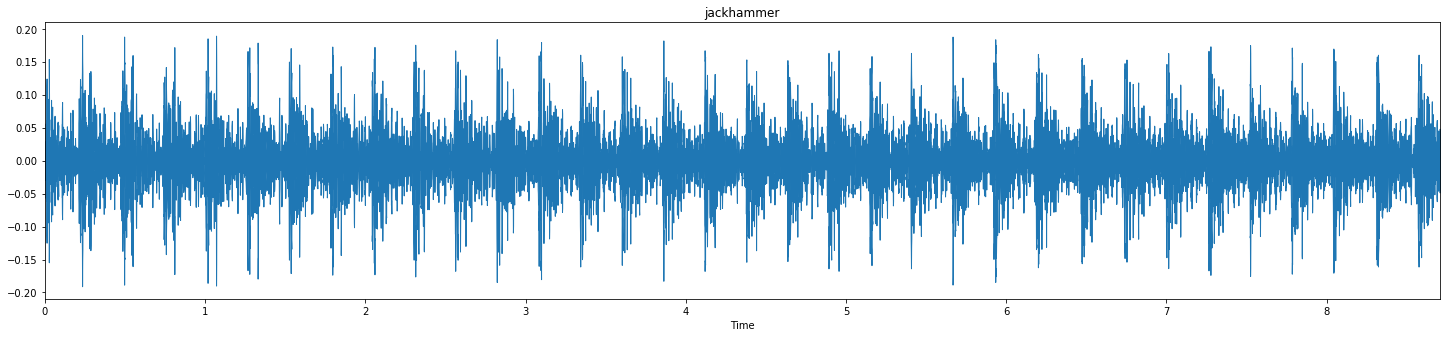
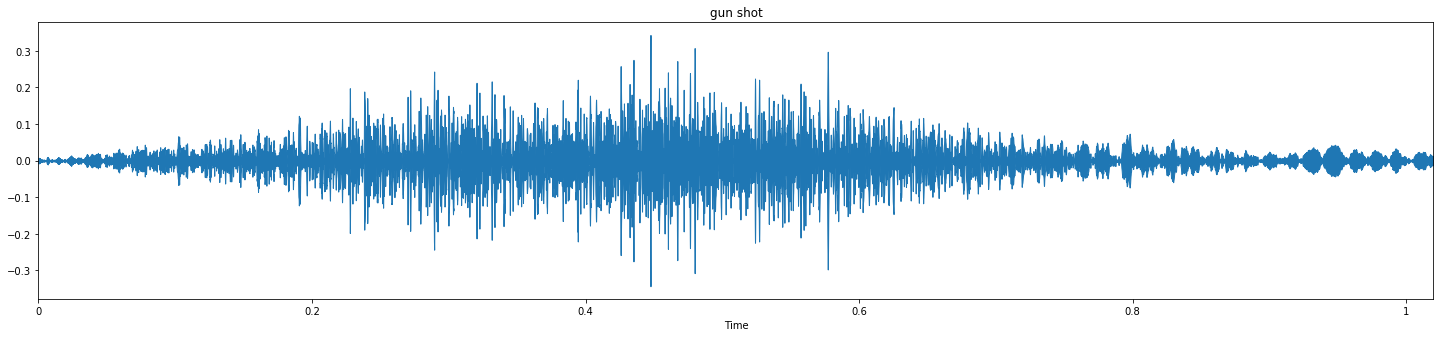
Sound sources will vary their compression and rarefaction in time significantly and it is easy to infer that based on the origin of the sound the time domain fluctuations of the sound signal will have unique signatures that can be used to classify the sound based solely on the distribution of amplitudes of the sound wave in time.

Below are the time domain signals for the 10 classes of sound which are to be classified









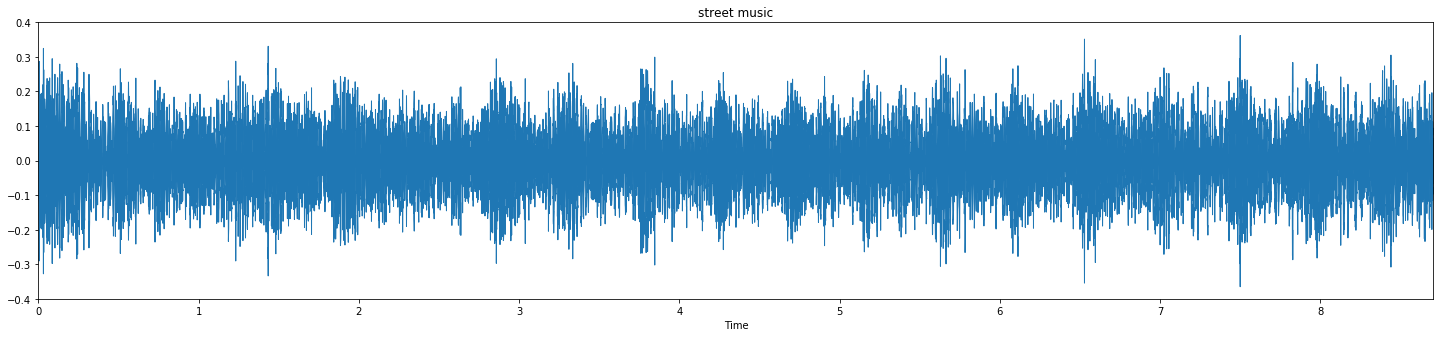
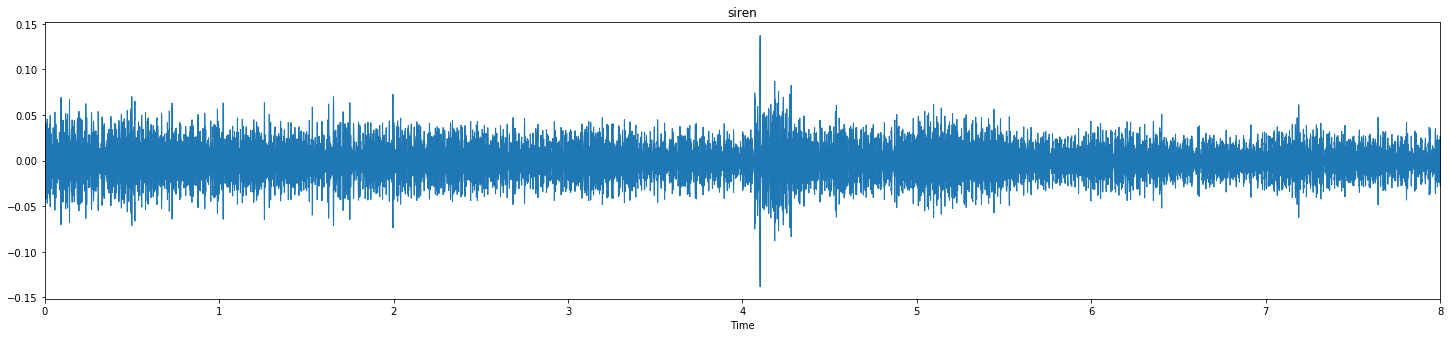


Figure 4. Time domain representation of sound classes.

1.4.2.1 Time Domain Histogram

**1.4.3 Feature Extraction (Frequency Domain)**

The Fourier transform decomposes a signal in our case sound into its various frequency components. Any signal can be considered to be a linear combination of frequencies weighted by coefficients that represent the strength of those frequencies over time.

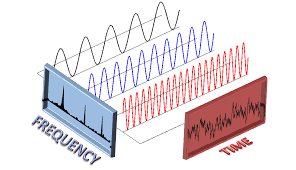


Figure 5. Decomposition of time domain signal into frequency domain

More formally any function can be represented as a weighted linear combination of sinusoidal and cosinusoidal functions over a period T Eq.1 . The coefficients of this weighted function can be determined by finding the coefficients ck by finding the projection of the initial function onto a set of orthonormal bases in the form of complex exponentials Eq.2.

(1)

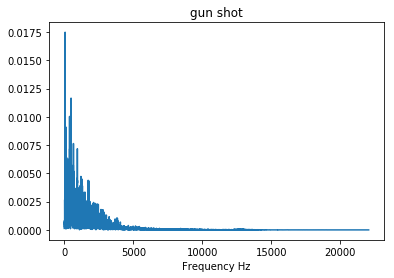
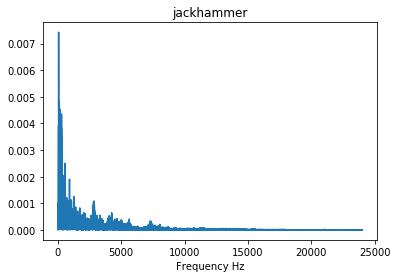
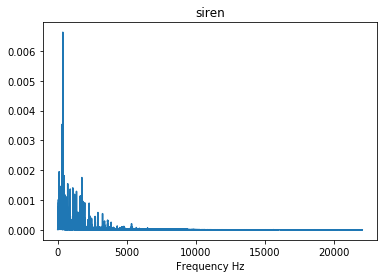
The projection of function f(t) on to the orthonormal basis determining the coefficients

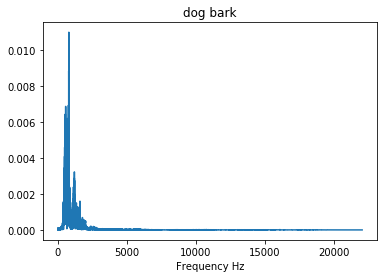
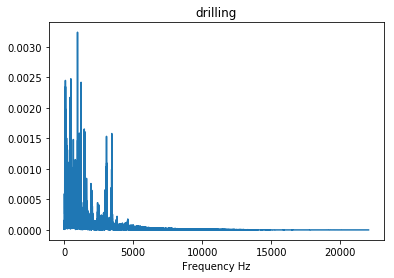
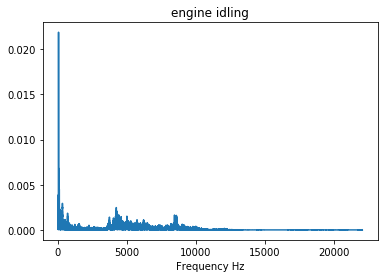
(2)

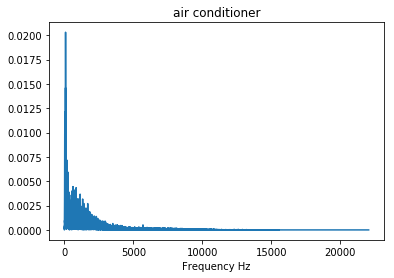
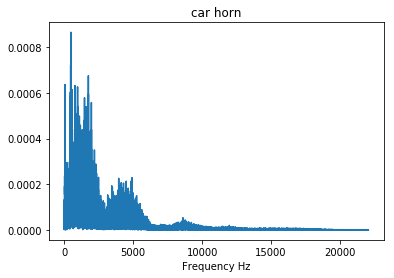
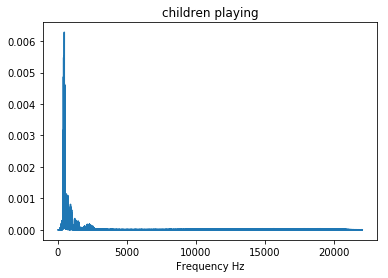
Taking the limit as T -> ∞ results in a continuous mapping from t space (time) into k space or frequency.

(3)

The frequency content of the 10 urban sound classes also represent vastly different frequency content that can be used as features for the input into machine learning algorithms. The sampling frequency of the sound signals is 20 KHz. Each of the training data samples is







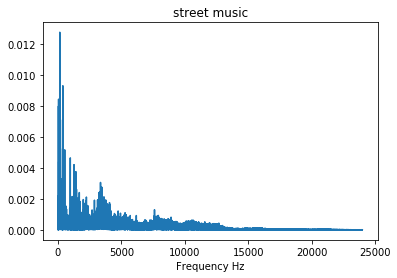
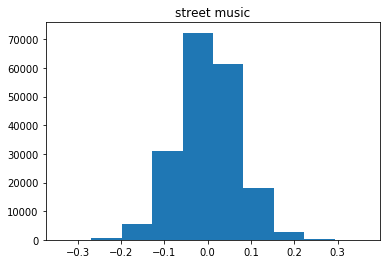
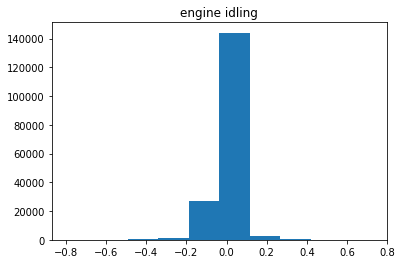
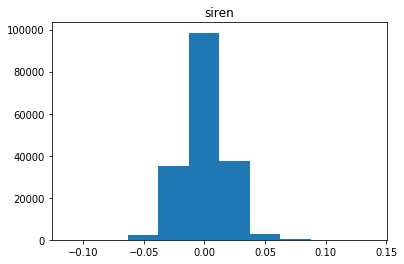
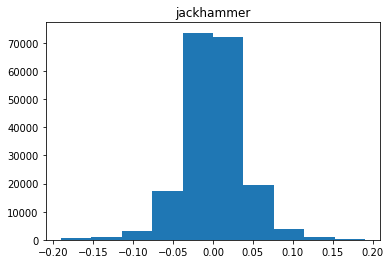
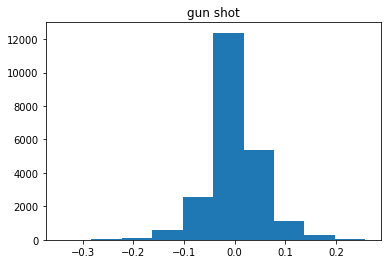
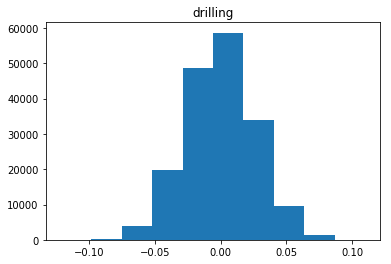
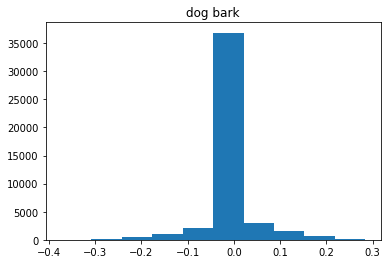
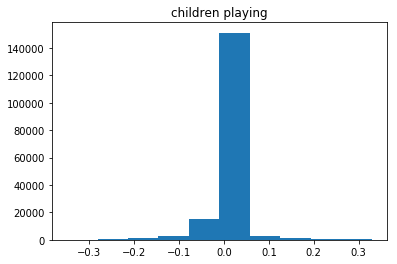
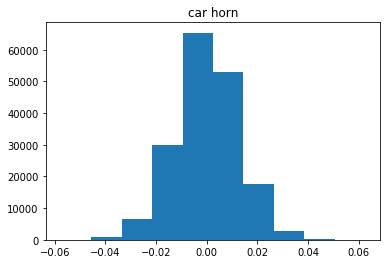
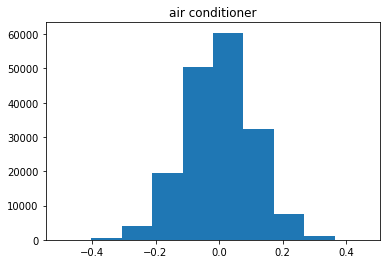


Figure 6. Frequency Domain Representation

**Distribution Characteristics**

As the sound signals can be characterized using their distributions it is in theory possible to classify the various classes using their distributions. This fact can be used to perform hypothesis tests (t-tests) to determine the likelihood of one class belonging to a particular distribution. Such classification would be dependent on whether the distributions of the signals are sufficiently different. The representation of the counts of signal values should be centered about zero due to the periodicity of the signals. Below is shown the distributions of all 10 classes which are centered about zero mean and a wide range of standard deviations.



Performing a t-test on one representative sample of each class results in the following data when comparing the air conditioner class to all other classes. If a significance level of .001 is chosen we can see from the table below that (upon a cursory examination) the classes can be separated by their sample statistics as observing difference of means this large given we assume the distributions to be the same are extremely unlikely.

|  |  |  |
| --- | --- | --- |
|  | Mean Difference | p-value |
| Air Conditioner | 0 | 1 |
| Car Horn | 9.904277188 | 4.38E-23 |
| Engine Idling | 2.53448109 | 0.011270508 |
| Gunshot | 2.806934777 | 0.005007206 |
| Jack Hammer | 16.1356969 | 2.81E-58 |
| Dog Barking | 2.092257551 | 0.036432201 |
| Siren | 19.79710196 | 2.28E-86 |
| Drilling | 4.043204537 | 5.29E-05 |
| Children Playing | 8.457763296 | 2.85E-17 |
| Street Music | 7.549893556 | 4.50E-14 |

**1.4.4 Feature Extraction (MFCC Spectrogram)**

Cepstral features allows the decomposition of the signal according to the so-called source-filter model widely used to model speech production. The signal is then decomposed into a carrier (the source, for speech it can be the glottal excitation) and a modulation (the filter, for speech it includes the vocal tract and the position of the tongue). MFCC’s are the most common cepstral coefficients [23]. They are obtained as the inverse discrete cosine transform of

the log energy in mel frequency bands. The filter bank representation of perceptual shift is shown below in Fig 7.

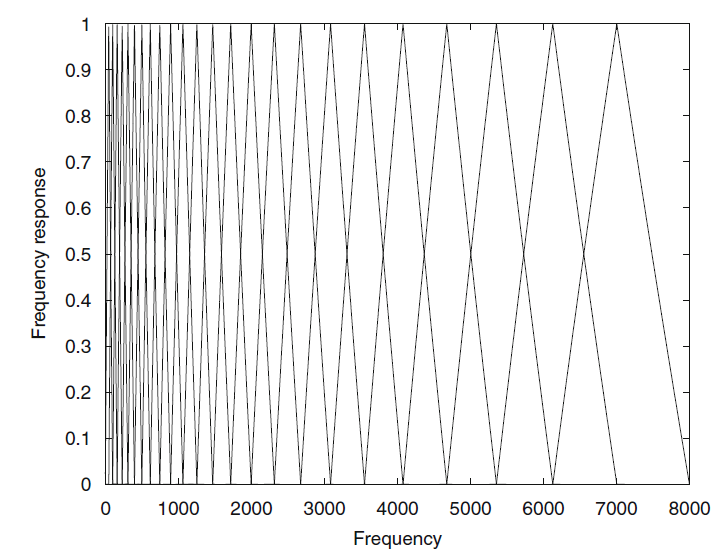
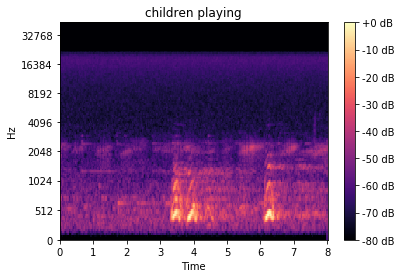
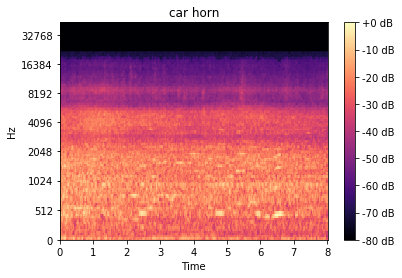
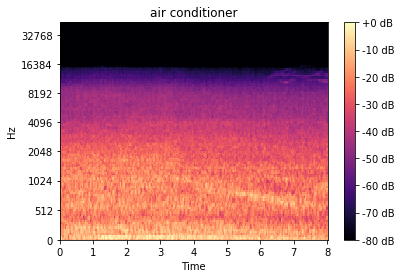
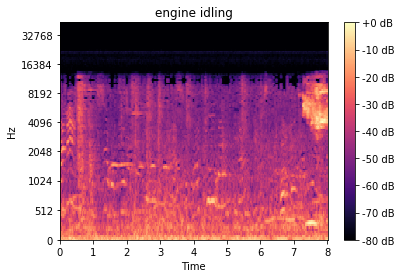
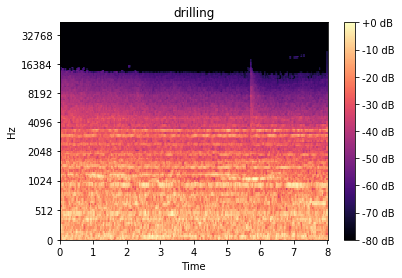
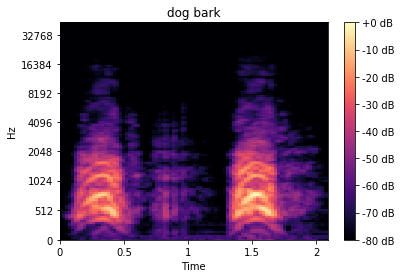
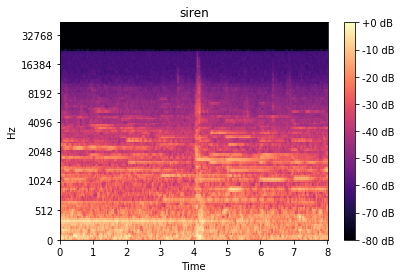
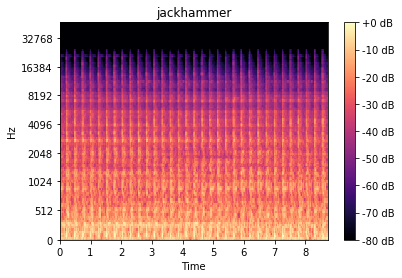
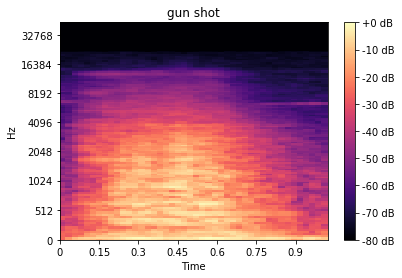


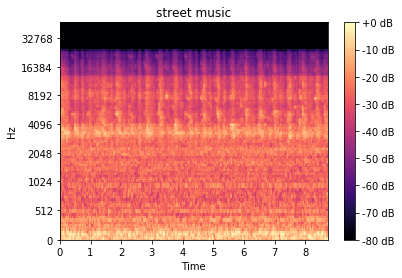
Figure 7. Mel Cepstral Perceptual Filter Bank

Applying the cepstral filter bank and obtaining the spectrogram of the cepstral coefficients for the 10 sound classifications is shown in Fig 8.



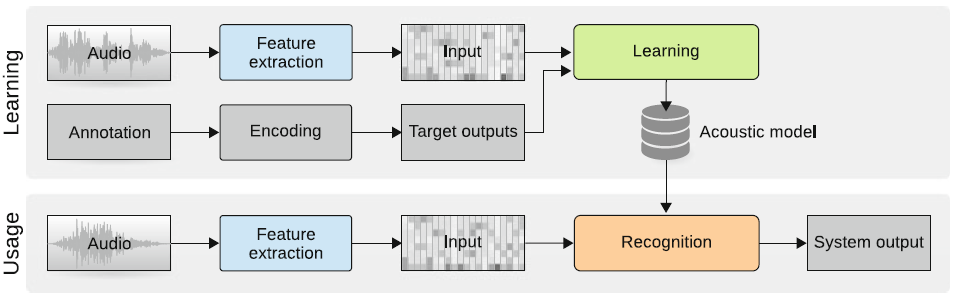






**1.5 Supervised Learning**

After the data acquisition and feature extraction steps, acoustic features and reference annotations for each audio signal are available. The next step is to *learn* a mapping between these features and class labels for sound classes, where the labels are determined from the reference annotations. This is based on a computational algorithm that can analyze and learn the similarities/differences between acoustic features and the class labels for various sound classes. The learned acoustic model is then used to assign a class label for acoustic features without reference annotations in the usage stage. The general framework for training a supervised model is given in Figure 9 below



**2. Convolutional Neural Networks**

There are two key ideas behind convolutional networks:

1. statistically meaningful interactions tend to concentrate locally, e.g., within a

short time window around an event;

2. shift-invariance (e.g., in time) can be exploited to share weights, thereby reducing

the number of parameters in the model.

Convolutional networks are well-suited to applications in which the desired output is a sequence of predictions, e.g., time-varying event detection, and the concepts being modeled derive only from local interactions.

Two-dimensional convolutional architectures are especially common, due to their natural application in image processing [6], which can in turn be adapted to time-frequency representations of audio. The benefits of two-dimensional convolutional architectures on time-frequency representations include a larger effective observation window (due to temporal framing, as one-dimensional convolutional networks), the ability to leverage frequency decompositions to more clearly locate structures of interest, and the potential for learning representations which are invariant to frequency transposition. Below is shown an example of a vanilla convolutional network and how it operates on two dimensional data. The audio data in this report is decomposed by 2D convolutional filtering of the image, subsampled and pooled in order to do classification on the 2D spectrograms presented in section 1.4.



**3 Experimental Design**

The sound class data was segmented into their classes and frames of the data decomposed using MFCC. After the decomposition of the data into spectrograms of the MFCC’s a convolution neural network with (3x3) filters was used to filter the data followed by subsampling and max pooling. After the completion of filtering a dense neural network was used in the classification process. The network was trained for 10 epochs. The following shows the results for accuracy for each epoch of training.

Epoch 1/10

200/200 [==============================] - 44s 218ms/step - loss: 9.2162 - acc: 0.0600 - val\_loss: 5.7885 - val\_acc: 0.0950

Epoch 2/10

200/200 [==============================] - 42s 208ms/step - loss: 4.3805 - acc: 0.2750 - val\_loss: 3.7967 - val\_acc: 0.1100

Epoch 3/10

200/200 [==============================] - 41s 205ms/step - loss: 2.5752 - acc: 0.5950 - val\_loss: 3.2991 - val\_acc: 0.1000

Epoch 4/10

200/200 [==============================] - 42s 210ms/step - loss: 1.9145 - acc: 0.7250 - val\_loss: 4.1190 - val\_acc: 0.1250

Epoch 5/10

200/200 [==============================] - 46s 231ms/step - loss: 1.5245 - acc: 0.7650 - val\_loss: 5.3967 - val\_acc: 0.1150

Epoch 6/10

200/200 [==============================] - 50s 248ms/step - loss: 1.4148 - acc: 0.7800 - val\_loss: 5.3986 - val\_acc: 0.0900

Epoch 7/10

200/200 [==============================] - 48s 242ms/step - loss: 1.1727 - acc: 0.8200 - val\_loss: 5.7613 - val\_acc: 0.0850

Epoch 8/10

200/200 [==============================] - 45s 223ms/step - loss: 1.1537 - acc: 0.8000 - val\_loss: 6.3192 - val\_acc: 0.0750

Epoch 9/10

200/200 [==============================] - 47s 237ms/step - loss: 1.1478 - acc: 0.8100 - val\_loss: 7.0458 - val\_acc: 0.0950

Epoch 10/10

200/200 [==============================] - 47s 234ms/step - loss: 1.2477 - acc: 0.8050 - val\_loss: 7.3352 - val\_acc: 0.1100

Final results show that based on CNN classification this method was able to obtain an 80% accuracy using a training set of 200 classes with 20 representation for each class.

References

[1] Giordano, B.L., McAdams, S., Rocchesso, D.: Integration of acoustical information in the perception of impacted sound sources: the role of information accuracy and exploitability. J. Exp. Psychol. Hum. Percept. Perform. **36**(2), 462–476 (2010). doi:10.1037/a0018388

[2] Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, New York (2006)

[3] Oppenheim, A.V., Schafer, R.W., Buck, J.R.: Discrete-Time Signal Processing. Prentice Hall, Upper Saddle River, NJ (1999)

[4] Brown, J.C.: Calculation of a constant q spectral transform. J. Acoust. Soc. Am. **89**(1), 425– 434 (1991)

[5] Tzanetakis, G., Essl, G., Cook, P.R.: Audio analysis using the discrete wavelet transform.In: Proceedings of the WSES International Conference Acoustics and Music: Theory and Applications (AMTA 2001), Skiathos, pp. 318–323 (2001)

[6] Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: SimpleMKL. J. Mach. Learn. Res. **9**, 2491–2521 (2008)