Spatiotemporal Analysis of Urban Noise Data

Ashwin Bhaskar Srivatsa, Sasanka Mouli Veleti and Sidharth Veluvolu

1 Introduction

Noise Pollution is a grievous problem which needs to be tackled, as it's a growing concern for many urban residents. Sounds that are not particularly loud but are nonetheless undesirable and uncontrollable can have serious implications for the listener, particularly if they occur over a long period of time [8]. The noise becomes much more upsetting if the source of noise is an agent or agency that has shown little concern for the individual who is suffering from the noise's effects and, as a result, nothing has been done to reduce the noise. Noise is not only inconvenient and annoying, but it has also been proven to be a health hazard, many have reported that they suffered with behavioral and emotional consequences, such as difficulty in sleeping, relaxing and feeling annoyed, angry or upset [3], and when intrusive noises continue, the body responds physiologically, and there is a risk of irreversible bodily damage - damage to the circulatory, cardiovascular, and gastrointestinal systems - over time [5].

In order to identify the source and mitigate this problem, there is a need to understand sound event detection [2]. Sound event detection is defined as recognition of individual sound events in audio, e.g., "dog barking, engine exhaust noise" requiring estimation of onset and offset for distinct sound as for identification of sound. There are multiple challenges for sound event detection and classifying them into various classes especially when it is to detect in multiple environments or large sets of data, or when there is overlapping of sound events or mixture of sound signals from sensors or identifying the similar sounds compared to the other distinct sound that is harder to classify the source [9]. So, it's important to evaluate the data set with actual recordings from urban noise sensors and to identify the sounds by classifying them into classes or labels. In this article we take the large data set which comprises of 3068 labeled 10 sec recordings from the Sounds of New York City (SONYC) acoustic network [1] (An acoustic network is a method of positioning equipment using sound waves) . Using this data, we plan to develop a machine learning model that classifies the audio sounds into particular categories based on the attributes or features and also further analyze it to find out the source. We also plan to address the mismatch of the testing data in this data set.

This can be done by applying machine learning algorithm on the data set and then analyzing it for the development of machine learning systems for real world urban noise monitoring. The timeline and work flow (Fig. 2.) for this is depicted below where The model that is devised from the data can be used on urban neighborhoods to better understand noise in that location and help the authorities mitigate this issue.

2 Related work

Collecting audio data and annotating the soundscapes are an important part of acoustic research especially in the situations with high variability in different locations.

In the paper [7] they have discussed about soundscapes and different ways to annotate audio data using crowdsourcing.There are basically two ways to achieve this one of them being waveform and the other one being spectrogram visualization. Certain annotation trials were done using the two techniques and interesting results were drawn from the experiments.People using the spectrogram visualization technique were able to produce high quality and precision annotations than waveform visualization.

According to [2] a supervised learning methodology is applied to real-life high quality recordings of 3-5 minutes with very little noise of 15 different acoustic scenes (lakeside beach, bus, cafe/restaurant, car, city center, forest path, grocery store, home, library, metro station,office, urban park, residential area, train, and tram) and two common environment areas (outdoor - residential areas and indoor - home). A Mel frequency Cepstral coefficient (MFCC) and Gaussian mixture model (GMM) was trained using expectation maximization algorithm. The overall accuracy achieved by the model is 72.5 % ranging from 13.9% for parks to 98.6% for office spaces. With our system we plan to include cross-validation to train our model so that there is no data contamination between train and test sets.

According to the work done in [6], audio annotation is critical for creating machine-listening systems, but there is little research on how to get accurate and timely crowd sourcing audio annotations. They aimed to quantify the reliability/redundancy trade-off in crowd sourced soundscape annotation, look at how visualizations affect accuracy and efficiency, and characterize how performance varies as a function of audio characteristics.

Our application will be based on the research paper's data [6], which presents us the process to collect acoustic data. SONYC has developed an acoustic sensor with high quality and low production cost to monitor the noise pollution levels across the city in neighborhoods like Manhattan, Brooklyn and Queens. The collected data was then annotated using a campaign on Zooniverse. The sensors follow DCASE (Detection and classification of acoustic scenes and events) to eliminate discrepancy. There are various other datasets like UrbanSound, UrbanSound8k that address this particular problem but have limited spacial and temporal data points.A VGGish model has been developed and trained using stochastic gradient descent to minimize cross-entropy loss. To eliminate over-fitting early stopping on validation set has been implemented. Two models were trained on course-level and fine-level tags.The overall AUPRC achieved by this model is 0.62 and 0.76 on different level classes, which performed poorly on music and non-machinery impact sounds. We plan to use this model and analyze the mismatches caused by the prediction on actual test data and find out the causes which led to these mismatches.

Over the years there has been a lot of research on annotation of audio data in different scenarios and predicting the source of noise in big cities. We will be extending this to analyze the mismatch of such predictions by leveraging machine learning metrics and coming up with our own model to predict the results with higher accuracy. A big part of our research will also be visualizing the locations of these mismatches. However most of our work focuses on analyzing existing models for the (SONYC-UST) dataset [1].

3 Data description

In order to build a system which would visualize the various sound points in and around a particular geographical location and also allow us to analyse the mismatches between the test and machine data, we need to have a dataset which has a diverse distribution of labeled sounds with spatial and temporal attributes. For which we have taken a dataset containing a training subset (13538 recordings from 35 sensors), validation subset (4308 recordings from 9 sensors), and a test subset (669 recordings from 48 sensors). Each recording has been annotated using a set of 23 "sound tags" like "engine presence, machinery presence,

- *Ashwin Bhaskar Srivatsa, E-mail: asriva36@uic.edu*
- *Sasanka Mouli Veleti, E-mail: svelet2@uic.edu*
- *Sidharth Veluvolu, Email: sveluv2@uic.edu*

non-machinery-impact presence, dog-barking-whining presence, music presence etc." [6].

```
fine:
    1:
        1: small-sounding-engine
        2: medium-sounding-engine
        3: large-sounding-engine
        X: engine-of-uncertain-size
    2:
        1: rock-drill
        2: jackhammer
        3: hoe-ram
        4: pile-driver
        X: other-unknown-impact-machinery
    3:
        1: non-machinery-impact
    4:
        1: chainsaw
        2: small-medium-rotating-saw
        3: large-rotating-saw
        X: other-unknown-powered-saw
    5:
        1: car-horn
        2: car-alarm
        3: siren
        4: reverse-beeper
        X: other-unknown-alert-signal
    6:
        1: stationary-music
        2: mobile-music
        3: ice-cream-truck
        X: music-from-uncertain-source
    7:
        1: person-or-small-group-talking
        2: person-or-small-group-shouting
        3: large-crowd
        4: amplified-speech
        X: other-unknown-human-voice
    8:
        1: dog-barking-whining
coarse:
    1: engine
    2: machinery-impact
    3: non-machinery-impact
    4: powered-saw
    5: alert-signal
    6: music
    7: human-voice
    8: dog
```

Fig. 1. A snapshot of the dataset labels

fig 1 shows all the fine level and course level classes. The course level lists down all the different catagories of sound samples that are captured, whereas fine level describes the all the sub-classes in these main classes. For example the group Engine has small-sounding-engine, medium-sounding-engine, large-sounding-engine as the fine-level classes.

The training, validation, and test subsets of the annotation data are contained in annotations.csv (see Figure 1). Each row in the file represents one multi-label annotation of a recording—it might be a single citizen science volunteer's annotation, a single SONYC team member's annotation, or the SONYC team's agreed-upon ground truth (for more information, see the annotator id in column description).The audio files used were recorded using the SONYC acoustic sensor network for monitoring urban noise pollution. In New York City, over 60 distinct sensors have been placed accumulating the equivalent of more than 50 years of audio data, of which a small fraction is used. The data is sampled by picking the closest neighbors based on VGGish qualities of recordings with recognized classes of interest. All of the recordings are 10 seconds long and were made with the same microphones and gain settings.

The sensors in the test set will not disjoint from the training and validation subsets, but the test recordings are displaced in time, occurring after any of the recordings in the training and validation subset. We plan to use the test data to find out the aggregate of mismatch by using a Multi Label classification Machine Learning Model like support vector machines and artificial neural networks.

In our work we will utilize the latitude and longitude columns to spatially locate the sound origin, the year, week, day, hour columns to precisely point the time and the category column which contain engine, machinery-impact, non-machinery-impact, powered-saw etc.to get the sound type to analyze and visualize the data.

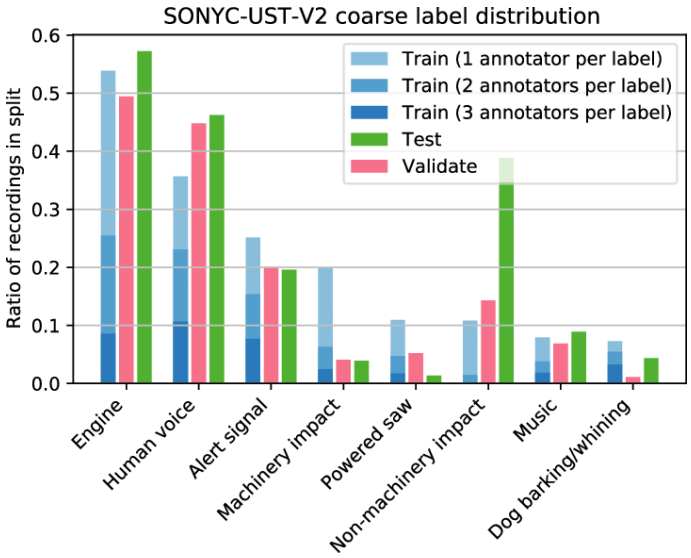

Fig. 2. Distribution of various sound tags in test, train, validate sets of dataset

4 RESEARCH

We propose to build a tool which will visualize the spatiotemporal values of the dataset, visuliaze the output of Machine Learning model and also visualize the points where the mismatches of testing and machine data occur based on the results of the Machine Learning model. We will be using Multi Label Classification Machine Learning Algorithms like support vector machines and existing neural networks to classify the audio files into various categorical sounds. Based on the results of ML model, we will dwell into understanding the mismatches which occur between the annotated and machine predicted data and analyze the causes behind it.

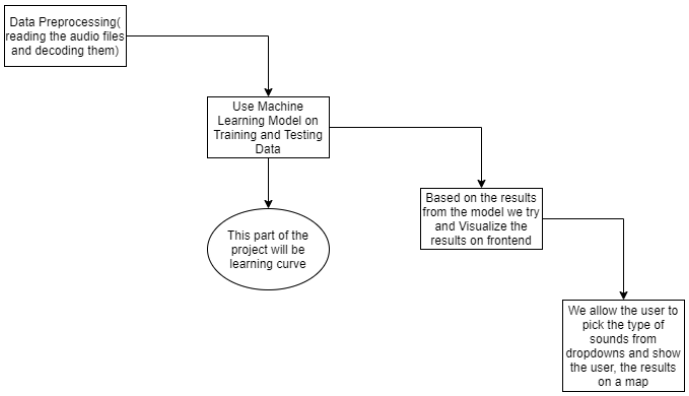

Fig. 3. Work flow diagram

We have divided the process of building this tool into four stages(see Figure 2) namely data preprocessing state where we process the audio files, Modeling stage where we apply the classification models on the dataset and create REST API's to interact with the front-end, Visualization stage where we visualize the results of machine learning model on

the front-end, component building stage where we build components which allows the user to pick spatial and temporal values of various sound samples and display the results on a map.

This tool would be useful to authorities who monitor sounds around the city which will allow them to pinpoint the location of sound origin.

Initially we started by collecting the dataset for this module which is [1]. This is a multi label classification dataset from the urban acoustic sensor network in the city of New York. The next step is to extract the embeddings from the audio data which is collected from 60 different sensors across the city of New York.To do this we implemented an algorithm called Open L3 which is a competitive and deep audio embedding based on the supervised L3-Net.OpenL3 is an improved version of L3-Net, and outperforms VGGish and SoundNet (and the original L3-Net) on several sound recognition tasks.

Next we used a MLP (Multi Layer Perceptron) using a single output layer of size 128 with ReLU activation function and using AutoPool to perfom some pooling functions. We trained the model using SGD (stochastic gradient descent) to reduce binary cross entropy loss. We also used L2-regularization to reduce the weights by adding a penalty term to the loss function.

SGD(stochastic gradient descent) is an iterative method for optimizing an objective function to best fit the model between the actual and predicted output.The main here is to move opposite to the gradient function and update the parameters until we reach the global minima. The learning rate determines the size of steps we take to reach the global minima. We used a learning rate of 0.001 and then with 0.1, we later used an adaptive learning rate optimizer like RMSProp optimizer. L2 regularization can deal with multicollinearity problems by constricting the weights in the function. It basically deals with adding L2 penalty which is square of the magnitude of coefficients.

Initially we trained the model on fine level class labels and generated the predictions. Similarly we trained the model on course level labels and generated the corresponding predictions.

We followed certain metrics to evaluate the performance of our model. The AUPRC (area under precision recall curve). The AUC curve is plotted with precision against recall with precision on y-axis and recall on x-axis. The area under the AUC curve basically explains the performance of the model against positive examples. This is relative to the number of positive examples in the problem.

```
Fine level evaluation:
======================
 * Micro AUPRC:            0.732747794782529
 * Micro F1-score (@0.5): 0.6185873605947956
 * Macro AUPRC:            0.5245790233574581
 * Coarse Tag AUPRC:
      - engine: 0.6451345992170708
      - machinery-impact: 0.5044753445334218
      - non-machinery-impact: 0.44699687079919237
      - powered-saw: 0.5153245680991896
      - alert-signal: 0.8216649672136981
      - music: 0.30422093048705173
      - human-voice: 0.9074674241587014
      - dog: 0.05134748235133917
```

Fig. 4. fine level predictions

As we can see from fig 3 the model has misclassified certain fine level tags leading to a lower AUPRC of 0.73 on fine level predictions as compared to 0.83 on course levcel predictions.The f1 score recorded for the fine level predictions is 0.61 and 0.73 for course level predictions. As we can observe here the model clearly performed poorly on fine level predictions. This is due to the dense heirarchy of class labels and the model failed to capture all the variance in the data.

We wanted to understand the mismatches caused by the model in these fine-level classes. We performed certain analysis on the predictions by this model on test and validation data. We started with combining the predictions and ground truth for each audio file by using pandas. This made it easy to perform various tests and analysis on our data.

5 RESULTS

When we plotted the sensors on map visually, we could see that most of the sensors are located in and around lower Manhattan city and others are positioned at upper Manhattan, Brooklyn City ((Fig. 12.). The sensors are located around Washington Square Park where most of the sounds are produced in an around the park.

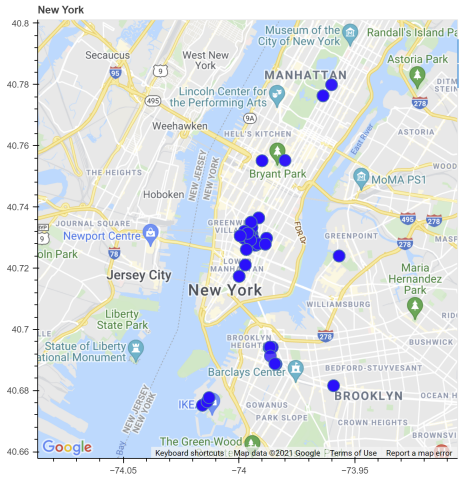

Fig. 5. Sensors located on the map

Based on the figure (Fig. 7.), the ground truth is visualized against the predicted values of the data set from the sensors. We can observe that against the sounds in ground truth, the predictions are consistent. For instance, the sensors which detect the sound 6-1 stationary-music, the predicted sounds are mostly person-or-small-group-talking or car-horn or dog-barking. We can assume that people who play stationary music don't prefer to play in loud sounds like large-sounding-engine.

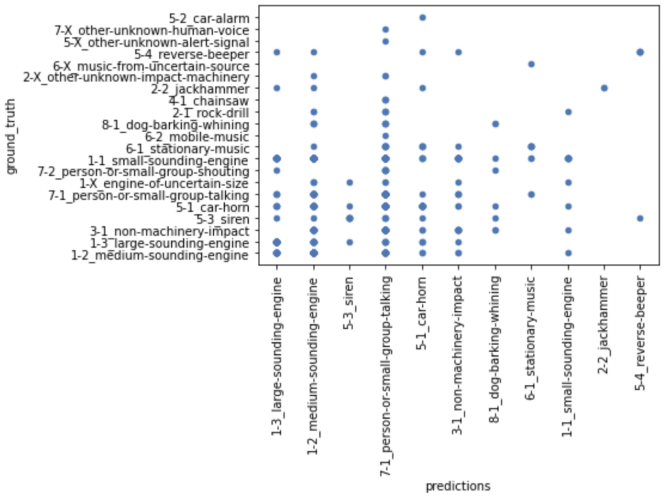

Fig. 6. Predictions vs ground truth

The mismatches detected from the sensors tells us that there seem to be more occurring on Fridays, week days compared to weekends (Fig. 8.). Least number of mismatches occurs on Sundays, we can estimate that less people would be working compared to other days. Similarly, more number of mismatches occur in morning and night hours (Fig. 9.). which we could assume that there could be more people heading to and returning from work and less mismatches between 12 and 16 hours.

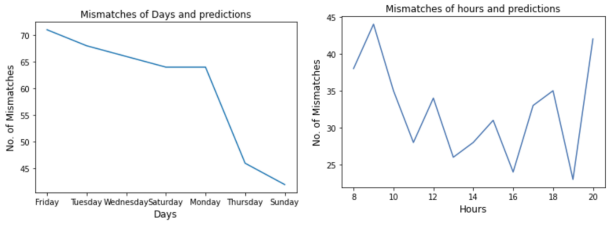

Fig. 7. No. of mismatches vs Days, Hours

The different sensors located in and around the city recorded many sounds and after calculating the mismatches from them through ML model, we could see that most of the mismatches occur in 28th, the least in 53th sensor ((Fig. 10.). The reason could be that more number of mismatches could occur where there is more sounds being produced, where it is less accurate to identify each particular sound, where as the other sensor could be located around a zone where there are less sounds recorded.

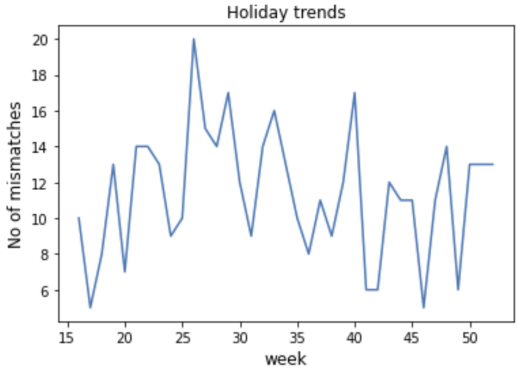

Fig. 8. No. of mismatches vs Sensors

When we further analyzed the data for any trends generated in holidays, there was an interesting pattern visible from the mismatches (Fig. 11.). The week of July 4th (week 27) saw an increase in the number of mismatches compared to rest of the year in 2019 which could be as a result of holiday weekend, more sounds are detected by the sensors across Manhattan city compared to rest of the year, there was an increase in number of mismatches. We could identify holidays in the year through sudden increase in the number of mismatches as seen the below figure.

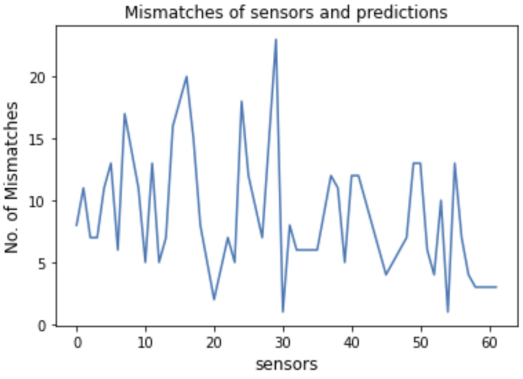

Fig. 9. Holiday trends through mismatches

Using the results obtained from our exploratory analysis, we developed a visual system which visualizes all the sensors points spatially and on a map. There are two components in the system. Namely, frontend and backend. The functionality of frontend is to contain all the temporal and spatial components required to understand mismatches of sounds for a given time and location for a particular sensor. The backend would serve all the API's required by frontend in visualize points of mismatches.

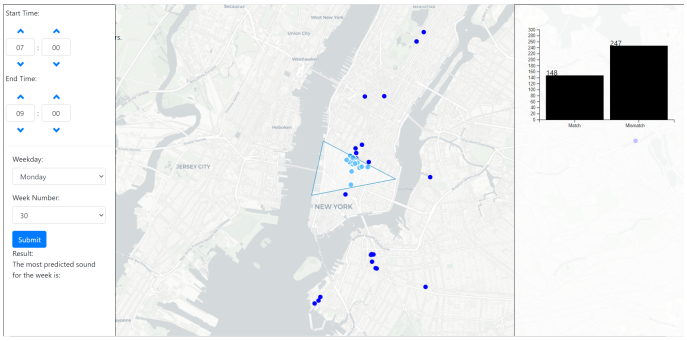

Fig. 10. Visual system visualizing number of mismatches

The (Fig 13.). shows a snippet of the developed visual system. For developing this system we have used Angular on the frontend and flask on the backend. The main advantage in using Angular is it provides us with component based architecture, this allows the developer segregate his functionality into components. Likewise we have segregated our functionality into 3 components, namely time component which allows the user to select the time frame of mismatches, map component which plots all the points of sensors and allows user interactions likes point click and polygon drawing and chart component which gives us with bar chart of matches and mismatches, predicted sound for a particular sensor or region. The main advantages of using flask is routing URL can be easily generated. For our system we have five API endpoints namely /sensors - fetches all the sensors spatial data, /particular:id- gets all the counts of mismatches and matches of a given sensor id, /soundpredicted - gets the most predicted sound for a particular region, /mismatcheschart - gets the number of matches and mismatches and get the most frequent sound, /mismatchestime - get mismatch data based on given time.

6 Conclusion

The main aim of this project is to understand and build a system to predict and visually analyze noise pollution in the city of New York. The dataset [1] is a multi-label classification dataset recorded by urban sensors across the city of New York. According to the analysis performed we can infer that the mismatches are more for fine level prediction as compared to coarse level prediction due to greater hierarchy of output classes in fine level taxonomy .By observing the model performance we could make out that it is biased towards certain sound groups in the city like small-sounding-engine and car-horn. This might be due to the fact that these are easily captured and can be mistaken for other sound groups like stationary-music and siren. According to our analysis the senors present in the Greenwich village right in the center of NYC have the most number of mismatches. The overlap of sounds in this part of the city during peak hours could be one of the reasons our model could not capture the relevant audio features.

There could be some tweaks that can be performed on this model to include data from various other regions and capture different sound groups in the city to reduce bias. This system can be expanded to capture noise data in other metropolis like Chicago, Los Angeles and San Francisco. This could also be further extended to include data from a range of years rather than concentrating on the data of 2019. Different type of neural networks and ML models can be used to achieve a higher F1 score. Nanocubes can be implemented to create interactive visualizations of large data points [4].

REFERENCES

[1] Sonyc urban sound tagging (sonyc-ust): a multilabel dataset from an urban acoustic sensor network.

[2] T. V. Annamaria Mesaros, Toni Heittola. Tut database for acoustic scene classification and sound event detection.

[3] A. Bronzaft. Neighborhood noise and its consequences.

[4] J. T. K. Lauro Lins and C. Scheidegger. Nanocubes for real-time exploration of spatiotemporal datasets, January 2013.

[5] T. K. S. M. S. Hammer and R. L. Neitzel. Environmental noise pollution in the united states: developing an effective public health response, 2013.

[6] J. C. V. L. G. D. H.-H. W. J. S. O. N. Mark Cartwright, Ana Elisa Mendez Mendez1 and J. P. Bello. Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network. Detection and Classification of Acoustic Scenes and Events 201, October 2019.

[7] J. S. A. W. S. M. D. M. E. L. J. P. B. Mark Cartwright, Ayanna Seals and O. Nov. Seeing sound: Investigating the effects of visualizations and complexity on crowdsourced audio annotations.

[8] W. H. Organization. Burden of disease from environmental noise: Quantification of healthy life years lost in europe, 2001.

[9] A. E. T. V. Toni Heittola, Annamaria Mesaros. Context-dependent sound event detection. EURASIP Journal on Audio, Speech, and Music Processing, January 2013.