# OPEN-WINDOW: A SOUND EVENT DATASET FOR WINDOW STATE DETECTION AND RECOGNITION

*Saeid Safavi*[1*], *Turab Iqbal*[1], *Wenwu Wang*[1], *Philip Coleman*[2], *Mark D. Plumbley*[1]

[1] Centre for Vision, Speech and Signal Processing, University of Surrey, UK,
{s.safavi, t.iqbal, w.wang, m.plumbley}@surrey.ac.uk
[2] The Institute of Sound Recording, University of Surrey, U.K.
p.d.coleman@surrey.ac.uk

## ABSTRACT

Situated in the domain of urban sound scene classification by humans and machines, this research is the first step towards mapping urban noise pollution experienced indoors and finding ways to reduce its negative impact in peoples' homes. We have recorded a sound dataset, called Open-Window, which contains recordings from three different locations and four different window states; two stationary states (open and close) and two transitional states (open to close and close to open). We have then built our machine recognition baselines for different scenarios (open set versus closed set) using a deep learning framework. The human listening test is also performed to be able to compare the human and machine performance for detecting the window state just using the acoustic cues. Our experimental results reveal that when using a simple machine baseline system, humans and machines are achieving similar average performance for closed set experiments.

***Index Terms***— dataset, sound event, deep neural network

## 1. INTRODUCTION

The acoustic distinction between outdoor and indoor scenes is an active research field and can be automated with some success [1, 2]. A much subtler difference is the change in the indoor soundscape induced by an open window. Being able to determine this, however, would allow applications in warning systems and be a prerequisite for an app-based urban sound mapping research.

Acoustic detection requires neither line of sight nor sensors at the window frame or knowledge of the number of windows or their size. The task, however, varies substantially in difficulty with the amount of sound inside and outside. From the point of machine classification, the lack of specificity is the most problematic aspect: very few sounds if any can be assumed to originate exclusively from outside and be present at all times to aid automatic detection. The required generalisation ability, however, can be assumed for humans, who might also use very subtle cues in the change of reverberations [2]. Since by changing the status of the window the acoustic characteristics of the room is changing, different features can be used as an input for machine learning methods to build accurate models, e.g. RT60, clarity, and reverberance. These features have been previously used by authors to automatically predict the perceived level of reverberation when there is no prior information about the room characteristics and results in promising outcome [3, 4].

To facilitate the study in this area, we have created a dataset which could be used to determine the degree of reliability with which an open window can be recognised by humans and machines under varying circumstances based only on acoustic cues. Since noise pollution is an increasing threat to the well-being and public health of city inhabitants [5], the recorded dataset can be used to investigate whether the findings for humans and machines can inform each other and can be used for further application-related research, e.g., active noise control applications to a larger region of control, such as in open windows and openings of noise sources [6].

## 2. THE DATASET

### 2.1. Overview

Audio data contained in this dataset is stored in WAV format and it is accessible from the Zenodo repository via this DOI: `10.5281/zenodo.3620748`.

### 2.2. Recording locations

The recordings have been made at three different locations.

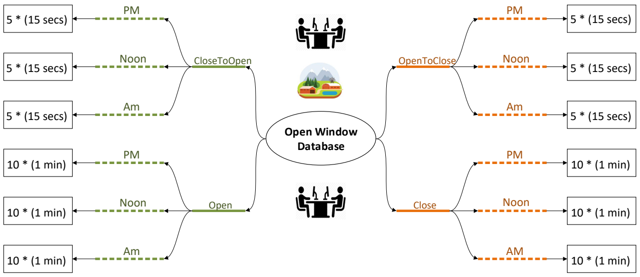

Figure 1: An overall schematic of the recorded dataset at three locations.

- Farm: A farm in Brook, Surrey, United Kingdom. The recordings were made in an open-plan studio flat area in the centre of the farm. The recordings in this location have the lowest levels of background noise, due mainly to a quiet environmental surrounding.
- Office 1: An office at the University of Surrey, Guildford, United Kingdom. The recordings were made in an open-plan

[*] This work was partially supported by the "Audio-Visual Media Research Platform" with grant reference number EP/P022529/1.

office located on the first floor, at the Centre for Vision, Speech and Signal Processing (CVSSP). Since this office accommodates 16 researchers, recordings in this location have the highest level of background noise.

- Office 2: An office at the University of Surrey, Guildford, United Kingdom. The recordings were made in a small size open-plan office at the CVSSP. This office accommodates 8 researchers and the recordings made in this office considered to have a medium level of background noise.

Figure 2.2 shows the schematic of the number of recordings made at each location. A plan view of the Office 2 is provided in Figure 2.

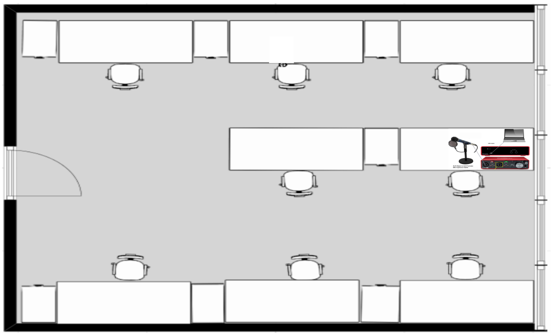

Figure 2: Plan view of the recording setup at Office 2.

### 2.3. Recording equipment

The recordings made at the two offices and a studio flat in a farm used a dedicated laptop, Focusrite Clarett 4pre USB external sound card (44,100 Hz sample rate at 16 bits per sample) [1], and a Behringer ECM 8000 microphone [2].

### 2.4. Recording setup

The Behringer ECM 8000 microphone is connected to the External Line Return (XLR) input of the Focusrite Clarett external sound card via an XLR cable. The external sound card is connected to the dedicated laptop and controlled using Ableton Live 10 [3] software for setting configurations and exporting the recorded audio files.

The microphone is located approximately 10 cm away from the window and fixed using a microphone holder. Figure 2.4 shows the photo of an actual setup in Office 1.

At each location 90 audio sessions are recorded; 60 one minute recordings for static state setup and 30 fifteen seconds recordings for transitional state setup.

### 2.5. File naming conventions

The naming convention for audio recording is as follows:

[Location]_[State]_[Time]_[IDX]

[State] will be one of the following: "O stands for open, C stands for Close, OC means a transition from Open to Close and CO stands for a transition from Close to Open." [Time] stamp will be one of the following: "AM stands for morning between 9:00 to 12:00, N stands for noon which is between 13:00 to 15:00 and PM which stands for an afternoon which is between 17:00 to 20:00." [IDX] is representing the file ID number.

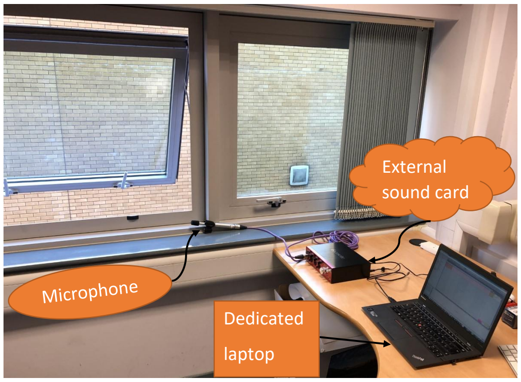

Figure 3: Image is taken from the recording setup at Office 1.

Table 1: Architecture of the baseline. 'Conv' stands for convolution, 'BN' stands for batch normalisation, 'Pool' stands for max pooling, and 'FC' stands for fully-connected.

| Layer | Output shape |
|---|---|
| | $640 \times 64 \times 1$ |
| Conv+BN+ReLU+Pool | $320 \times 32 \times 32$ |
| Conv+BN+ReLU+Pool | $160 \times 16 \times 64$ |
| Conv+BN+ReLU+Pool | $80 \times 8 \times 128$ |
| Conv+BN+ReLU+Pool | $40 \times 4 \times 256$ |
| Conv+BN+ReLU+Pool | $20 \times 2 \times 512$ |
| Global Average Pool | 512 |
| FC+Softmax | 4 |

For example, "Farm_C_PM_01.wav", means this file is recorded at the farm and in the afternoon when the window is closed and the file ID is 01.

## 3. MACHINE RECOGNITION BASELINE

In this section, we present the baseline system that we developed for the Open-Window dataset. The source code for this baseline is freely available[4]. Our aim was to develop a simple system that is of relatively low complexity, yet powerful enough to achieve good results. To this end, our baseline is a six-layer convolutional neural network (CNN) and the features used are logarithmic mel-spectrograms (log-mels). We first describe the feature extraction stage that produces the log-mels. Following this, we describe CNN in more detail, including the architecture and the training procedure. Finally, we describe how we split the data for training and evaluation.

[1] https://focusrite.com/usb-c-audio-interface/clarett-usb/clarett-4pre-usb

[2] https://thomann.de/gb/behringer_ecm_8000.htm

[3] https://ableton.com/en/live/

[4] https://github.com/saeidsafavi/OpenWindow

### 3.1. Feature Extraction

The Open-Window dataset provides recordings as 16-bit PCM waveforms sampled at 44.1 kHz. Instead of using the waveforms directly, our baseline system extracts log-mel spectrograms and then uses these spectrograms as inputs for the CNN. The log-mels were extracted by taking the short-time Fourier transform, squaring the result, scaling the frequency axis using mel filter banks, and finally scaling the magnitude using a logarithmic function. We used a window length of 2048 samples (46 ms), a hop length of 1032 samples (23 ms), and 64 mel bins.

In the Open-Window dataset, the audio clips vary in length. The fixed-state classes are approximately 60 s in length, while the transition classes are approximately 15 s. To ensure the inputs of the neural network are fixed in length, our system partitions the log-mels into blocks with shape $640 \times 64$, which corresponds to 15 s of audio. Audio clips that are less than 15 s are padded with zeros, while clips that are longer but not evenly divisible are padded with zeros if the remainder is greater than 10 s and truncated otherwise.

### 3.2. Convolutional Neural Network

The CNN that we use is comprised of 5 convolutional layers and a single fully-connected layer. Each convolutional layer is followed by batch normalisation [7], a ReLU activation function [8], and $2 \times 2$ max-pooling in that order. The number of output features for each convolutional layer is detailed in Table 1. After the final convolutional layer, the spatial dimensions are reduced to a scalar by taking the average. The resulting 512 features are then mapped to $K = 4$ class probabilities using a fully-connected layer and a softmax non-linearity. The total number of parameters is slightly over 3.5 M. As a result of this, the computational demands and memory requirements are relatively low.

The models are trained using the categorical cross-entropy loss function, which is defined as

$$L(y_i, \hat{y}_i) = -\sum_{i}^{K} y_i \log(\hat{y}_i), \tag{1}$$

where $y$ is a one-hot vector representing the ground truth and $\hat{y}$ is the output of the CNN. For optimisation, our baseline uses the Adam algorithm [9] with a learning rate of 0.0005. The learning rate is decayed by 10 % after every two epochs to help with convergence. Training is carried out for 50 epochs with a batch size of 64. After each epoch, the model state is saved and the validation set accuracy is recorded. The final model used for inference is the one that achieved the highest validation set accuracy.

### 3.3. Training set split

To train and evaluate the baseline, the Open-Window dataset was split into three subsets: a training set, a validation set, and a test set. As a means to do this, the dataset was first split into six folds, which were generated in a way that balances the number of instances from each class and location across folds. Initially splitting the dataset into folds allows more flexibility in how they can be used for the training/validation/test split. In our case, we used fold 1 for the test set, fold 2 for the validation set, and folds 3-6 for the training set. This means the training set contains two thirds of the recordings, or approximately 2.25 hours of audio. The exact mapping of instances to folds is available as part of the dataset release. This is so that fair comparisons can be made with the baseline in future work.

Table 2: Experimental results for the baseline system. The accuracy and the mean average precision (mAP) are reported along with 95% confidence intervals.

| System | Accuracy | mAP |
|---|---|---|
| ClosedSet | 79.2%±1.83% | 87.7%±1.18% |
| ClosedSet-50% | 60.2%±3.87% | 65.1%±3.07% |
| OpenSet-O1/O2/F | 35.4%±3.02% | 44.4%±3.15% |
| OpenSet-O1/F/O2 | 82.9%±2.39% | 88.2%±1.52% |
| OpenSet-F/O1/O2 | 41.3%±1.74% | 42.8%±0.96% |

## 4. BASELINE RESULTS

To evaluate the baseline, we looked at two different scenarios:

- Closed-set scenario: The training set, validation set, and test set are sampled from the same distribution. More specifically, they all contain audio clips from all three of the recording locations.
- Open-set scenario: The training set, validation set, and test set contain audio clips from different recording locations.

For the closed-set scenario, we used the training/validation/test set split described in Section 3.3; fold 1 is used for the test set, fold 2 is used for the validation set, while the remaining folds are used for the training set. For the open-set scenario, we look at three different configurations for the training/validation/test splits:

- Office 1/Office 2/Farm (OpenSet-O1/O2/F)
- Office 1/Farm/Office 2 (OpenSet-O1/F/O2)
- Farm/Office 1/Office 2 (OpenSet-F/O1/O2)

These configurations result in approximately 90 training clips. Using the folds as in the closed-set scenario, there are approximately 180 clips in the training set. To be able to make a fair comparison between the two scenarios, we also present results for the baseline when trained on fold 5 and fold 6 only, and refer to the system as ClosedSet-50% in the tabulated results.

To score the systems, we used two metrics: accuracy and mean average precision (mAP). The accuracy is the percentage of correctly classified instances. The mAP is defined as

$$\mathsf{mAP} = \frac{\sum_{k=1}^{K} \mathsf{AP}_k}{K},$$

where $\mathsf{AP}_k$ is the average precision for class $k$. The average precision is roughly the area under the precision-recall curve for class $k$. To account for random variation, we ran ten trials for each system and averaged the scores. 95% confidence intervals are also provided.

The results are presented in Table 2. Comparing the results from the closed-set experiments, it can be seen that halving the number of training clips from 180 to 90 reduces the accuracy by almost 20 % in absolute terms and the mAP by more than 20 %. This demonstrates the importance of training data and that a subpar amount of data can lead to poor performance. Observing the open-set results, it can be seen that leaving out clips from the Farm location during training and testing can improve the performance greatly; the accuracy increases from 60.2 % to 82.9 %. Moreover, only testing clips from the Farm location or training exclusively with clips from the Farm location drastically decreases the performance; the accuracy decreases by more than 20 %. It should be noted, however, that the accuracy is still notably higher than 25 %, which is the expected accuracy for

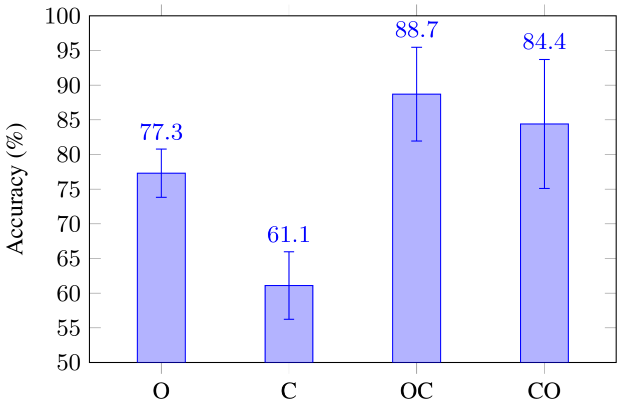

Figure 4: Class-wise performance of the ClosedSet system.

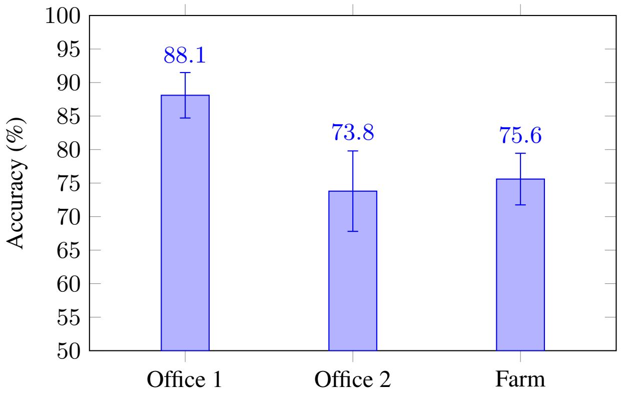

Figure 5: Location-wise performance of the ClosedSet system.

random guessing. This low performance may be attributed to the Farm location specifically or to the fact that it is a very different location in general, which demonstrates how difficult it can be to recognise sounds from unseen environments. As we will show next, this latter reason for low performance is more likely.

To analyse the ClosedSet system in more detail, Figure 4 presents the class-wise performance of the system. It can be seen that the transition classes are the easiest to classify, followed by the Open state class. The system did poorly on the Closed state class, which is understandable considering there is a less acoustic activity during this state. In Figure 4, we present the location-wise performance of the system. It can be seen that recordings in Office 1 are easiest to classify, followed by Farm and Office 2 in that order. These results show that the Farm location is not an inherently difficult location for the baseline despite the lack of acoustic activity in this location.

## 5. HUMAN LISTENING TEST

Human window state classification performance for fixed and transitional states are also measured, using the same test utterances from the automatic classification experiments. Eight listeners, mainly research fellows at the CVSSP, participated in the evaluations. Each participant initially trained for two separate tests; using static states (by listening to nine randomly selected files, three audio files per location and per state) and transitional states (by listening to nine randomly selected files, three audio files per location and per state). Then each participant listened to 120 audio files for each of two separate experiments, for static state experiment each of duration approximately 60 seconds, and for transitional state experiment each of duration approximately 15 seconds, in a quiet room using the same computer and headphones. Since the human listener was trained using audio files from all locations then their obtained performances are comparable with the closed set results from the machine baseline.

Table 3: Performance of human listeners in identifying the window status. *Trans stands for transitional state experiment.

| Setup | Accuracy | Location |
|---|---|---|
| Static state | 68.33% | Average |
| Trans* state | 88.75% | Average |
| Static state | 61.25% | Farm |
| Static state | 78.75% | Office 1 |
| Static state | 65.00% | Office 2 |
| Trans* state | 75.00% | Farm |
| Trans* state | 96.25% | Office 1 |
| Trans* state | 95.00% | Office 2 |

### 5.1. Results

Table 3 shows window state classification performance achieved by human listeners for both static state and transitional state tests. The results are presented as an average per each setup and then for further analysis divided further per each location.

Table 3 shows that humans are very accurate in classifying transitional window states in a noisy environment (Office 1) by achieving an accuracy of 96.25%. Evidently, humans have difficulty classifying static window state in a quiet places (e.g. Farm), the average accuracy is 61.25%.

A comparison of average performances for two different setups, static and transitional states, reveals that humans can benefit from sequential information in the transitional state setup and classify more accurately for transitional states than static state events.

## 6. CONCLUSIONS

In this paper, Open-Window, which is a manually recorded audio dataset, is presented. Full description of this audio dataset is described and the data is released on Zenodo so it can be used for research purposes by others. The recorded dataset contains around 3.5 hours of audio data recorded at three locations and from two different recording setups; static and transitional state. Open-Window is the first audio dataset recorded that reflects the differences in the acoustic cues of indoor and outdoor environments.

This dataset is available for research and development in a variety of fields like security, and living enhancement. To validate the consistency of Open-Window, two baseline systems have been tested and compared; machine and human baseline. The obtained results show that humans outperform machine learning methods in classification of window state. The results suggest that including sequential features could improve the machine performance further for the transitional state setup. In this research, we have used log-mel spectrograms as an input for the CNN baseline. Other features can also be used as an input to build more accurate models, e.g. RT60, clarity, and reverberance.

## 7. REFERENCES

[1] M. Hornikx, “Acoustic modelling for indoor and outdoor spaces,” 2015.

[2] B. Locher, A. Piquerez, M. Habermacher, M. Ragettli, M. Röösli, M. Brink, C. Cajochen, D. Vienneau, M. Foraster, U. Müller, *et al.*, “Differences between outdoor and indoor sound levels for open, tilted, and closed windows,” *International journal of environmental research and public health*, vol. 15, no. 1, p. 149, 2018.

[3] S. Safavi, A. Pearce, W. Wang, and M. D. Plumbley, “Predicting the perceived level of reverberation using machine learning,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2018, pp. 27–30.

[4] S. Safavi, W. Wang, M. D. Plumbley, A. J. Choobbasti, and G. Fazekas, “Predicting the perceived level of reverberation using features from nonlinear auditory model,” in *Proceedings of the 23rd FRUCT conference*. Institute of Electrical and Electronics Engineers (IEEE), 2018, pp. 527–531.

[5] H. Ising, B. Kruppa, *et al.*, “Health effects caused by noise: evidence in the literature from the past 25 years,” *Noise and Health*, vol. 6, no. 22, p. 5, 2004.

[6] T. Murao, M. Nishimura, and W.-S. Gan, “A hybrid approach to active and passive noise control for open windows,” *Applied Acoustics*, vol. 155, pp. 338–345, 2019.

[7] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *International Conference on Machine Learning*, vol. 37, Lille, France, 2015, pp. 448–456.

[8] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ser. ICML’10, Madison, WI, USA, 2010, pp. 807–814.

[9] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations*, San Diego, CA, 2015.