ESC: Dataset for Environmental Sound Classification

Karol J. Piczak
Institute of Electronic Systems
Warsaw University of Technology
Warsaw, Poland
K.Piczak@stud.elka.pw.edu.pl

ABSTRACT

One of the obstacles in research activities concentrating on environmental sound classification is the scarcity of suitable and publicly available datasets. This paper tries to address that issue by presenting a new annotated collection of 2 000 short clips comprising 50 classes of various common sound events, and an abundant unified compilation of 250 000 unlabeled auditory excerpts extracted from recordings available through the Freesound project. The paper also provides an evaluation of human accuracy in classifying environmental sounds and compares it to the performance of selected baseline classifiers using features derived from mel-frequency cepstral coefficients and zero-crossing rate.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Content Analysis and Indexing; H.5.5 [**Information Systems**]: Sound and Music Computing

Keywords

environmental sound; classification; dataset

1. INTRODUCTION

Recent years have brought a steady stream of advances in machine perception. Computer systems undertake progressively more complex tasks, at times even surpassing human capabilities. A significant part of these spectacular achievements has come in visual recognition, with recent proliferation of successful deep learning approaches.

At the same time, research in auditory recognition tasks has been focusing mostly on speech and music processing. Analysis of environmental sounds (a very diverse group of everyday audio events which cannot be described as speech nor music) has lagged behind in applying those recent improvements, despite numerous possible applications in audio surveillance systems [9], hearing aids [2], smart room monitoring [16], and video content highlight generation [3].

MM'15, October 26–30, 2015, Brisbane, Australia.

ACM 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806390.

One of the objective impediments to more active research in this field is strong fragmentation and difficulty in comparability and reproducibility. Most studies so far (Barchiesi et al. [4] and Chachada & Kuo [5] present recent surveys of this topic) have been conducted on datasets that are either very specific, small, or (semi-)proprietary [1, 7, 8, 10, 11, 13, 15, 17]. This scarcity of publicly available datasets[1] and difficulty in accessing the original code for study replication make research reproducibility efforts harder than they should be. That is in stark contrast to such fields as computer vision where corpora like MNIST[2] and CIFAR[3] have been prominently used as a de facto standard for baseline comparisons. Only in the recent months did initiatives such as the Urban Sound project [14] (a dataset of recordings concentrating on urban environments) bring some hope for a change in this matter. The situation, however, still remains rather bleak.

Therefore, the goal of this paper is to facilitate open research in the field of environmental sound classification by:

- contributing a publicly available dataset of environmental recordings,
- presenting estimates of human classification accuracy for this dataset,
- comparing these numbers with baseline performance of most common machine learning classifiers,
- providing a Jupyter (IPython) notebook with a more thorough analysis and code for easy replication of obtained results.

2. THE ESC DATASET

The presented compilation consists of three parts: the main labeled set comprising 50 classes of various environmental sounds, a small proof-of-concept subset of 10 classes selected from the main dataset - serving as a simplified benchmark - and a supplementary dataset of unlabeled excerpts suitable for unsupervised learning experiments.

All datasets consist of sound clips constructed from recordings available publicly through the Freesound project [6]. Classes included in the labeled part of the dataset were arbitrarily selected with the goal of maintaining balance between major types of sound events, all the while taking into consideration the limitations in the number and diversity of available source recordings, and subjectively assessed usefulness and distinctiveness of each class. The Freesound database of field recordings was queried for common terms

[1] Most datasets are listed on a website maintained by Toni Heittola: http://www.cs.tut.fi/~heittolt/datasets.html [*Accessed Aug. 5, 2015*]
[2] http://yann.lecun.com/exdb/mnist/ [*Accessed Aug. 5, 2015*]
[3] http://www.cs.toronto.edu/~kriz/cifar.html [*Accessed Aug. 5, 2015*]

related to the constructed classes. Search results were individually evaluated and verified by the author by annotating fragments containing events belonging to the given class. These annotations were then used to extract 5-second-long recordings of audio events (shorter events were padded with silence as needed). The extracted samples were reconverted to a unified format (44.1 kHz, single channel, Ogg Vorbis compression at 192 kbit/s). The labeled datasets were consequently arranged into 5 uniformly sized cross-validation folds, ensuring that clips originating from the same initial source file are always contained in a single fold.

The resulting dataset is available under a Creative Commons non-commercial license through the Harvard Dataverse project[4]. It is accompanied by an IPython notebook containing a more thorough analysis of the dataset than is possible in a short paper format, and the detailed results obtained along with source code for study replication[5].

2.1 ESC-50

The *ESC-50* dataset consists of 2 000 labeled environmental recordings equally balanced between 50 classes (40 clips per class). For convenience, they are grouped in 5 loosely defined major categories (10 classes per category):

- animal sounds,
- natural soundscapes and water sounds,
- human (non-speech) sounds,
- interior/domestic sounds,
- exterior/urban noises.

The goal of the extraction process was to keep sound events exposed in the foreground with limited background noise when possible. However, field recordings are far from sterile, thus some clips may still exhibit auditory overlap in the background.

The dataset provides an exposure to a variety of sound sources - some very common (*laughter*, *cat meowing*, *dog barking*), some quite distinct (*glass breaking*, *brushing teeth*) and then some where the differences are more nuanced (*helicopter* and *airplane* noise).

One of the possible deficiencies of this dataset is the limited number of clips available per class. This is related to the high cost of manual annotation and extraction, and the decision to maintain strict balance between classes despite limited availability of recordings for more exotic types of sound events. Nevertheless, it will, hopefully, be useful in its current form and is a concept that could be expanded on if sufficient interest is expressed.

2.2 ESC-10

The *ESC-10* is a selection of 10 classes from the bigger dataset, representing three general groups of sounds:

- transient/percussive sounds, sometimes with very meaningful temporal patterns (*sneezing*, *dog barking*, *clock ticking*),
- sound events with strong harmonic content (*crying baby*, *crowing rooster*),
- more or less structured noise/soundscapes (*rain*, *sea waves*, *fire crackling*, *helicopter*, *chainsaw*).

This subset should provide an easier problem to start with, and it was initially constructed as a proof-of-concept dataset. The task of classifying sounds from such a constrained set of classes, a trivial feat from a human perspective, sets the bar really high for accuracy expected from automatic sound recognition systems. Therefore, this subset presents a slightly different problem to tackle than the whole *ESC-50* dataset. The differences between classes are much more pronounced, with limited ambiguity, and as such it may favor a different kind of machine learning approaches.

2.3 ESC-US

Unfortunately, the limited number of instances available in the labeled part of the dataset makes it rather inadequate for more complex knowledge discovery approaches like learning representations from data. To mitigate this issue, an additional dataset of 250 000 recordings (extracted from Freesound files tagged as "field recording") is provided in the same short-clip (5-second-long) format. It should be more fitting for procedures involving unsupervised pre-training and generative models.

Although the *ESC-US* dataset should be treated as not hand-annotated and is presented as such, it does include the metadata (tags/sound descriptions) submitted for the original recordings. However, in contrast to the labeled part of the dataset, the metadata were not verified individually by the author, but rely solely on Freesound's quality control procedures through crowd moderation.

Therefore, the dataset, apart from clustering and manifold learning experiments, could be also used in weakly supervised learning regimes (classification with labels partially missing or not specific enough).

3. SOUND CLASSIFICATION

3.1 Human classification accuracy

The human auditory system has little problem recognizing a plethora of sound stimuli, even in very noisy conditions. Therefore, it is to be expected that, with such a limited challenge as presented by the dataset, proper recognition of sound events should not be difficult at all. The real question was: how easy is it? To answer this, numerous participants were asked through the CrowdFlower crowdsourcing platform to try their best at classifying sounds from the labeled datasets.

The experiment involved presenting a number of sound recordings to participants and asking them to choose a correct label from a list of 10 or, respectively, 50 categories. The participants were paid a flat fee per unit of work (classifying 10 recordings). Quality control was maintained through internal CrowdFlower's procedures (participant pre-screening and ongoing monitoring through randomly inserted test questions with an expected answer). The final results were further assessed by the author for potential outliers, and a small number of entries were eliminated this way. In total around 4 000 judgments were collected for each dataset (on average a dozen human classification entries per individual clip of *ESC-10* and two for *ESC-50*). Although it is hard to come up with formal statistical interpretation with such an experiment setup, it should nevertheless provide a rough estimate of human capabilities in recognizing everyday sounds.

The average accuracy achieved was 95.7% for the *ESC-10* dataset and 81.3% for *ESC-50*. Recall for individual classes varied greatly between types of sound events - from 34.1% for washing machine noise to almost 100% for crying babies

[4] http://dx.doi.org/10.7910/DVN/YDEPUT
[5] https://github.com/karoldvl/paper-2015-esc-dataset

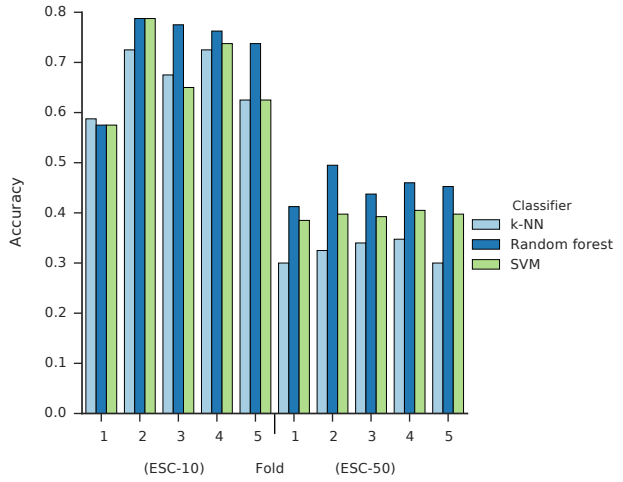

Figure 1: Comparison of classification accuracy between folds depending on the choice of classifier.

and barking dogs. Without going too much into details[6], sound events presented in the dataset can be divided into three groups based on their difficulty level:

- easy categories (most human sounds, animals, and some very distinct sound sources, like *siren*, *water drops*, *breaking glass*),
- average categories (everything in between the other two),
- difficult categories (mostly soundscapes and various mechanical noises).

One of the problems in such experiments is that with increasing number of categories, it becomes more difficult for untrained participants to mentally grasp all the possibilities and semantic differences. With 50 classes of sound events, it was still possible to provide them in one coherent view (divided into 5 major groups for faster orientation) without reaching for nested taxonomies, but it was on the verge of what is verifiable in such an experiment setup.

Nevertheless, based on these experiments, one can expect that trained and attentive listeners could score flawlessly on the smaller dataset and most probably achieve accuracy levels reaching 90% on the main dataset, with some room for error when classifying more ambiguous mechanical noises and soundscapes.

3.2 Baseline machine classification results

Having established an approximate figure on what is the desired target accuracy for a sound recognition system with near-human capacity, the second goal was to verify what can be achieved with some baseline approaches to machine classification of environmental sounds. The aim of this analysis was not to construct the most robust system possible, but to investigate what can be done with basic approaches, exploring potential pitfalls and intricacies of the dataset.

Two types of features were extracted from each clip: zero-crossing rate and mel-frequency cepstral coefficients (*MFCC*). The former is a very simple, yet useful feature, whereas the latter are ubiquitous in speech processing and analyzing

[6] Full experiment results are available as a spreadsheet supplementing the dataset, and a more thorough analysis is performed as part of the provided IPython notebook.

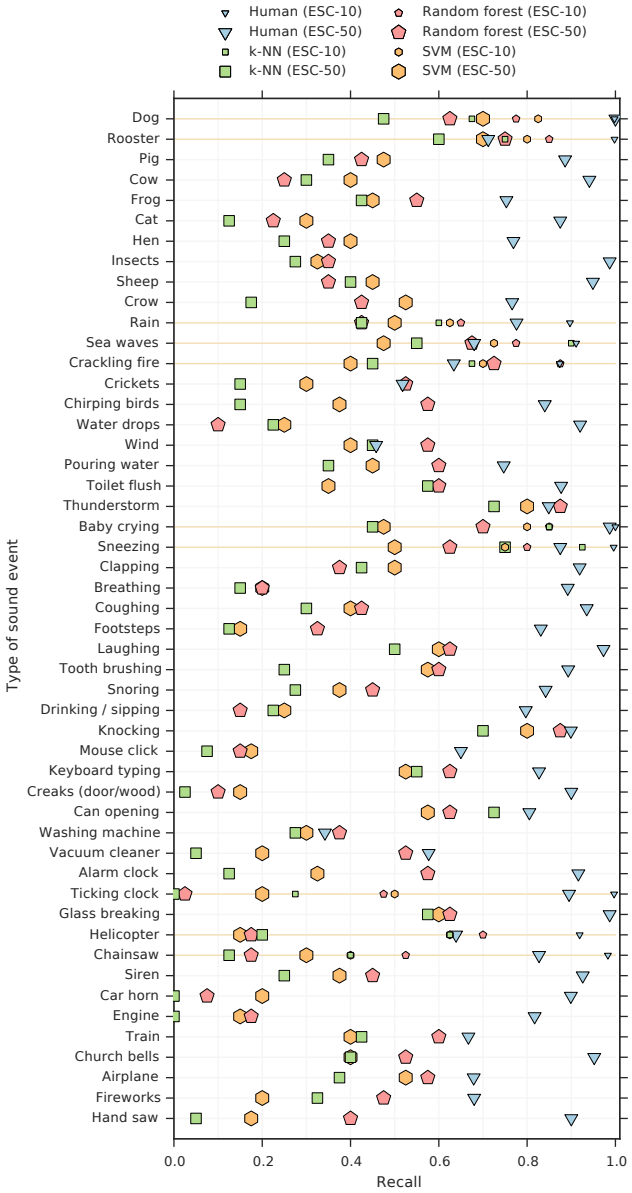

Figure 2: Category recall for different types of classifiers and human assessment.

harmonic content. MFCCs were computed using the librosa package[7] with default settings resulting in frame length of 11.6 ms. Discarding the 0th coefficient, first 12 MFCCs and zero-crossing rate were summarized for each clip with their mean and standard deviation across frames. Feature vectors created in this way were treated as input to three types of classifiers: k-nearest neighbors (*k-NN*), random forest ensemble and support vector machine (*SVM*) with linear kernel. Learning was performed on both datasets with a 5-fold cross-validation regime.

The *ESC-10* dataset had an average classification accuracy ranging from 66.7% for the k-NN classifier to 72.7% for the random forest ensemble, with SVM in the middle (67.5%). Some significant dispersion in results achieved could be seen

[7] *librosa: v0.3.1* library by B. McFee et al., DOI: http://dx.doi.org/10.5281/zenodo.12714 [*Accessed Aug. 5, 2015*]

between folds, owing to their small absolute size (see figure 1). The *ESC-50* dataset had less variability between folds when validating the models, but more pronounced outperformance by the random forest ensemble (44.3%) as compared to SVM (39.6%) or k-NN (32.2%).

One tendency that stands out when contrasting performance of different classifiers on both datasets is the pronounced drop in accuracy for the simplest (k-NN) model. It could indicate that the dependencies between features were more intricate in the bigger dataset, and they were better captured with more complex models.

In general, these rudimentary classification systems performed poorly when contrasted with their human counterparts, yet the difference is more pronounced for some groups of sounds than the others (see figure 2). For instance, lots of recordings in soundscape/background noise categories prove to be quite ambiguous for human listeners, a group which coincidentally scores quite high with automated systems.

The SVM classifier performed better for animal sounds than the random forest ensemble. Although a possible artifact of the data, it may also indicate that using more customized models for specific broader groups of sounds could be a viable option (creating a form of hierarchical multi-stage classification system).

It should be noted that the presented baseline classification methods are relatively simple. An evaluation of more robust approaches (based on convolutional neural networks) is performed in a more recent work of the author [12].

4. SUMMARY

The aim of this paper was to present a new compilation of environmental recordings that could enrich the research domain not so abundant with publicly available datasets. Hopefully, this material will help foster more open research efforts in analyzing environmental sounds.

There are numerous possible ways to expand on this topic, some of which include:

- compiling a fully replicable survey comparing various approaches utilized in past research papers,
- evaluating deep neural networks and other deep learning models in the context of environmental sound classification, e.g. with unsupervised pre-training using the *ESC-US* dataset,
- exploring the *ESC-US* dataset and the research possibilities it creates (clustering techniques, using available metadata in weakly-supervised setting or a hybrid machine-crowd annotation project, manifold learning etc.).

Acknowledgments

I would like to thank Frederic Font for his help in using the Freesound API and anonymous reviewers for their thorough and helpful comments.

5. REFERENCES

[1] BBC sound effects library. http://www.sound-ideas.com/sound-effects/bbc-sound-effects.html. (*Aug. 5, 2015*).

[2] E. Alexandre et al. Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2249–2256, 2007.

[3] L. Ballan et al. Deep networks for audio event classification in soccer videos. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 474–477, 2009.

[4] D. Barchiesi et al. Acoustic scene classification: Classifying environments from the sounds they produce. *Signal Processing Magazine*, 32(3):16–34, 2015.

[5] S. Chachada and C.-C. J. Kuo. Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, 3:e14, 2014.

[6] F. Font, G. Roma, and X. Serra. Freesound technical demo. In *Proceedings of the ACM International Conference on Multimedia*, pages 411–412. ACM, 2013.

[7] D. Giannoulis et al. Detection and classification of acoustic scenes and events: An IEEE AASP challenge. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2013.

[8] I. Lallemand, D. Schwarz, and T. Artieres. Content-based retrieval of environmental sounds by multiresolution analysis. In *Proceedings of the Sound and Music Computing conference*, 2012.

[9] K. Łopatka, P. Zwan, and A. Czyżewski. Dangerous sound event recognition using support vector machine classifiers. In *Advances in Multimedia and Network Information System Technologies*, pages 49–57. Springer, 2010.

[10] J. Maxime et al. Sound representation and classification benchmark for domestic robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6285–6292. IEEE, 2014.

[11] T. Nishiura and S. Nakamura. An evaluation of sound source identification with RWCP sound scene database in real acoustic environments. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, volume 2, pages 265–268. IEEE, 2002.

[12] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015. *In press.*

[13] A. Plinge et al. A bag-of-features approach to acoustic event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3704–3708. IEEE, 2014.

[14] J. Salamon, C. Jacoby, and J. P. Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the ACM International Conference on Multimedia*, pages 1041–1044. ACM, 2014.

[15] D. Stowell and M. D. Plumbley. An open dataset for research on audio field recording archives: freefield1010. *arXiv preprint arXiv:1309.5275*, 2013.

[16] M. Vacher, J.-F. Serignat, and S. Chaillol. Sound classification in a smart room environment: an approach using GMM and HMM methods. In *Proceedings of the IEEE Conference on Speech Technology and Human-Computer Dialogue*, pages 135–146, 2007.

[17] M. van Grootel, T. Andringa, and J. Krijnders. DARES-G1: Database of annotated real-world everyday sounds. In *Proceedings of the NAG/DAGA International Conference on Acoustics*, 2009.