GENERALISATION IN ENVIRONMENTAL SOUND CLASSIFICATION: THE ‘MAKING SENSE OF SOUNDS’ DATA SET AND CHALLENGE

Christian Kroos [1], Oliver Bones [2], Yin Cao [1], Lara Harris [2],
*Philip J. B. Jackson [1], William J. Davies [2], Wenwu Wang [1], Trevor J. Cox [2], Mark D. Plumbley [1] **

[1] University of Surrey
Centre for Vision, Speech and Signal Processing (CVSSP)
Guildford, Surrey, GU2 7XH, UK

[2] University of Salford
Acoustics Research Centre
Manchester M5 4WT, UK

ABSTRACT

Humans are able to identify a large number of environmental sounds and categorise them according to high-level semantic categories, e.g. urban sounds or music. They are also capable of generalising from past experience to new sounds when applying these categories. In this paper we report on the creation of a data set that is structured according to the top-level of a taxonomy derived from human judgements and the design of an associated machine learning challenge, in which strong generalisation abilities are required to be successful. We introduce a baseline classification system, a deep convolutional network, which showed strong performance with an average accuracy on the evaluation data of 80.8%. The result is discussed in the light of two alternative explanations: An unlikely accidental category bias in the sound recordings or a more plausible true acoustic grounding of the high-level categories.

***Index Terms*—** Acoustic classification, machine learning challenge, sound taxonomy, deep learning, convolutional neural network

1. INTRODUCTION

Management of audio data typically involves assigning textual descriptors and allocating audio to a predefined category. Previous novel approaches to the problem of organising audio data into categories include: Augmenting the WordNet framework [1, 2] with audio concepts in order to classify sounds [3, 4]; using Gaver’s [5] taxonomy based upon the mechanical properties of sound-causing events in an audio retrieval system [6]; classifying urban noise complaints [7]; classification by affect ratings [8]; and using hyponym generation from web text with subsequent manual refinement [9].

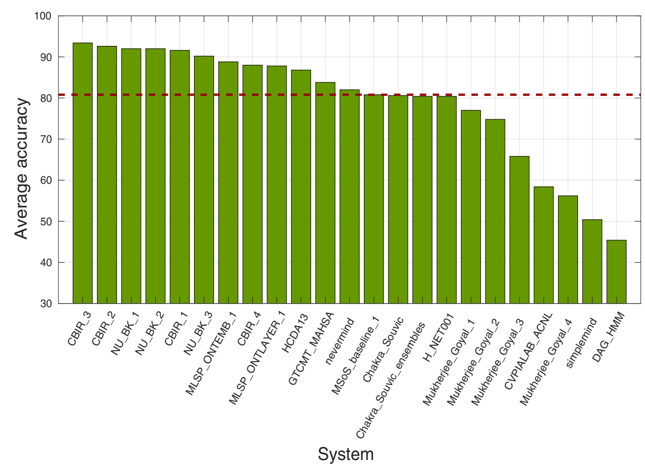

Fig. 1. Classifications results of the submitted systems with the baseline performance marked with the dashed red line.

The data set and taxonomy presented here constitutes a first approach to use empirical data obtained from human participants in a controlled experiment. It is limited in scope (only 60 basic starting terms were selected; see section 2), but it explores the highest level of abstraction that can be derived from human categorisations. Thus the resulting top level categories refer to broad concepts and cover a wide variety of sound-types that often seem to share little essential acoustic properties. Our interest from the view point of signal processing and machine learning was whether machine systems could replicate this top-level categorisation.

To encourage exploration of the topic, we created the ‘Making Sense of Sounds’ Data Challenge within the research context of the acoustic signal processing and machine learning project with the same name[1]. The challenge follows

*This work was supported by EPSRC grant EP/N014111/1 ‘Making Sense of Sounds’ and by European Commission H2020 research and innovation grant 688382 ‘AudioCommons’.

[1] http://cvssp.org/projects/making_sense_of_sounds

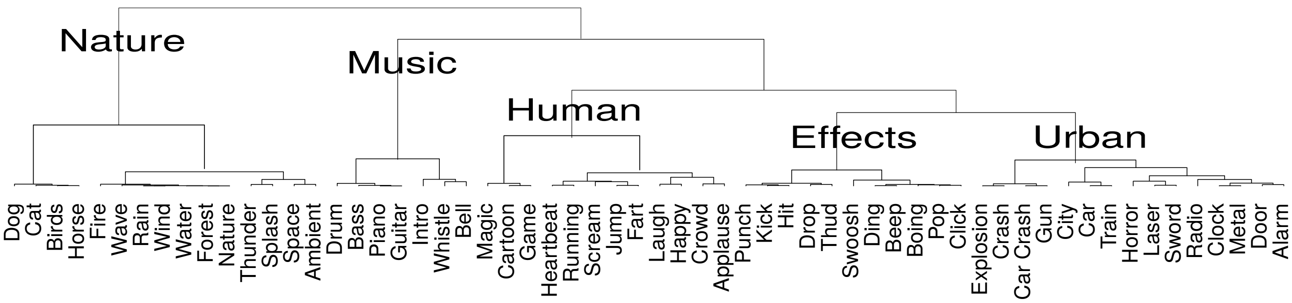

Fig. 2. The dendrogram with labelled clusters resulting from cluster analysis of the dimensions generated by correspondence analysis.

the tradition of previous machine learning challenges in the field, in particular, the Detection and Classification of Acoustic Scenes and Events challenge (DCASE 2013 [10], 2016 [11], 2017 [12] and 2018 [13]).

The current challenge differs from the DCASE tasks in focussing on a few very broad categories. The emphasis on semantic generalisation also distinguish it from challenges that focus on specific topic areas such as the Bird Audio Detection challenge [14] (detection of bird calls) or the MLSP 2013 Bird Classification Challenge [15] (acoustic bird species classification).

2. THE DATA SET

The categories of the data set presented here were derived from human experiments [16]. In brief, audio files corresponding to each of the top 60 search terms entered by users of Freesound[2] were downloaded from Freesound for use as experimental stimuli. The category label data spontaneously-generated by $N = 101$ participants during a sorting task were analysed using correspondence analysis and agglomerative hierarchical cluster analysis, using Wards criterion (see [17]), producing a dendrogram. Correspondence analysis is a method similar to principal component analysis but is suitable for categorical rather than continuous data (see [18, 19]). The dendrogram was sliced at the point at which the ratio of between-cluster inertia to total inertia was 0.1, creating five clusters (see Figure 2). This ratio was chosen to create enough labels so as to be meaningful without compromising the quality of the labelling. Each of the resulting five clusters was given a category name according to the category labels that were most over-represented in that cluster. Significance of over-representation of each descriptive word within each cluster was assessed using a hypergeometric distribution [19].

To create the data set, 2000 audio files were compiled by collecting 400 sound-types belonging to each of the five categories. Files were taken from three sources: the above mentioned Freesound data base, the ESC-50 data set [20] and the Cambridge-MT Multitrack Download Library[3]. The raw files were processed so as to have an identical format: Single-channel 44.1 kHz, 16-bit WAV files. File length was uniformly set to 5 seconds, but in some cases the target sound did not fill the entire duration and short periods of silence were included.

3. THE CHALLENGE

The aim of the challenge was to explore how machine learning systems would fare if they had to categorise sounds into categories determined by human judgement.

The five major categories (*Nature*, *Human*, *Music*, *Effects* and *Urban*) were the target classification labels. Within each class the provided training data consisted of varying sound-types, e.g., different animals in the *Nature* category or different instruments in the *Music* category such as *guitar* and *mandolin*. Most of the sound-types were, of course, represented by several instances themselves, but as a rule these instances originated from different recordings, e.g., different guitars recorded with different microphones in varying situations. The machine classifier was therefore forced to generalise well in order to be successful, something humans achieve seemingly effortless: Based upon previously established schemas, humans are capable of generalising from past experience to new sounds, e.g. recognising a dulcimer or a kora as a musical instrument despite having never heard this instrument before.

The data set was (pseudo-) randomly split per category in a development set (1500 sound clips) and a held-out evaluation set (500 sound clips). For the development set the category labels together with the sound-type labels as additional information were published. It was not guaranteed that the number of samples for each sound-type was proportionally the same in the development set and the evaluation set or even that a particular sound-type was represented at all in both data sets. The task for the challenge participants was to classify the audio files of the evaluation data set according to the five

[2] https://freesound.org/

[3] http://www.cambridge-mt.com/ms-mtk.htm

Layer	Feature map
log mel spectrogram	$T \times 64, 1$
convolutional layer $[3 \times 3, 64]$, BN, ReLU $[3 \times 3, 64]$, BN, ReLU	$T \times 64, 64$
2×2 max pooling	$T/2 \times 32, 64$
convolutional layer $[3 \times 3, 128]$, BN, ReLU $[3 \times 3, 128]$, BN, ReLU	$T/2 \times 32, 128$
2×2 max pooling	$T/4 \times 16, 128$
convolutional layer $[3 \times 3, 256]$, BN, ReLU $[3 \times 3, 256]$, BN, ReLU	$T/4 \times 16, 256$
2×2 max pooling	$T/8 \times 8, 256$
convolutional layer $[3 \times 3, 512]$, BN, ReLU $[3 \times 3, 512]$, BN, ReLU	$T/8 \times 8, 512$
2×2 max pooling	$T/16 \times 16, 512$
global max pooling $512 \times$ class number	$1 \times 1, 512$
class number fc, softmax $1 \times$ class number	1×1, class number
total parameters	$4,690,116$

Table 1. Configuration of the baseline network system

categories. Determining the sound-type was not required but admissible. This allowed for two major strategies:

1. Fine-grained classification on the sound-type level followed by an additional step that maps sound-types to categories.
2. Direct classification of the high-level categories.

As performance measure average accuracy was chosen:

$$A = \frac{1}{C} \sum_{c \in \mathbb{C}} \frac{n_c^{true}}{N_c} \tag{1}$$

where $\mathbb{C}$ is the set of categories and C its cardinality, N_c the number of sound clips belonging to category c and n_c^{true} the number of correct classifications with respect to class c.

4. BASELINE SYSTEM

We built a baseline system based on Convolutional Neural Networks.

As input features for the baseline system log mel-spectral coefficients were chosen, commonly used in supervised separation with neural network classifiers. The system itself is based on VGG model with 8 convolutional layers. A filter kernel size of 3×3 is used in the convolutional layers, followed by batch normalisation [21] to ensure the stability of the distribution of nonlinearity inputs. This reduces the chance that the optimiser gets stuck in a saturated regime, accelerating the training. A ReLU is then applied after the batch normalisation. Global max pooling (GMP) is utilised at the end of the last convolutional layer to summarise the feature maps to a vector. Finally, a fully-connected layer is applied to the summarised vector followed by a softmax nonlinearity. The probabilities of the audio classes are then generated. For the loss function cross-entropy was selected following standard procedures in multi-class problems. The detailed configuration of the network is shown in Table 1.

	Development	**Evaluation**
Effects	85.7	88.0
Human	84.1	81.0
Music	94.3	95.0
Nature	77.9	70.0
Urban	77.2	70.0
Average	83.8	80.8

Table 2. Average accuracy of the baseline system in percent

5. RESULTS

5.1. Baseline

A four-fold cross-validation was applied to the development set. To that end, the training data were randomly split into four folds, containing each 25% of the data. Three folds (75% of the data) were used for training, the remaining fold was used for validation. All four combinations of the folds were tested and the average precision computed. The results are shown in Table 2.

The baseline system was also tested on the evaluation set. The system was developed, however, strictly without reference to the evaluation data and only a single output of predicted class membership was evaluated in the same way ordinary entries are evaluated. The results are also depicted in Table 2. For a closer inspection of the classification error, the confusion matrix of the baseline system with regard to the evaluation data set is displayed in Figure 3.

5.2. Challenge contributions

Twenty-two systems from 11 teams were submitted, originating both from academia and industry and from a variety of countries (e.g., USA, India, France, Greece). The winning system achieved an average accuracy of 93%. The results for all systems including the baseline are shown in Figure 1. All systems with one notable exception were based on deep learning methods. The systems of five of the teams used transfer learning and the overwhelming majority of systems worked directly on the categories and did not consider the lower-level sound-types. More details can be found on the challenge website[4].

6. DISCUSSION

The baseline showed a strong performance with an average accuracy of 80.8% on the evaluation set. In particular, the category *Music* was very well distinguished from all other categories, achieving 95%. The highest errors are found in categories *Nature* and *Urban*, which both reach only 70% accuracy. However, the error is primarily not a mutual confusion: *Nature* is most frequently misclassified as *Urban*, but *Urban* is most often confused with *Human*.

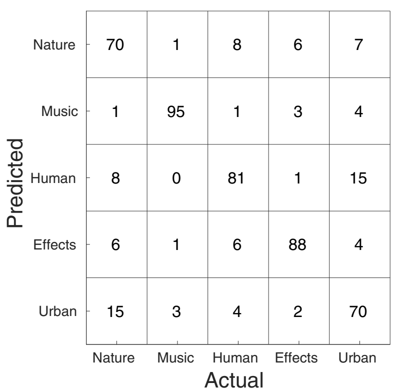

Fig. 3. Confusion matrix of the baseline classification on the evaluation set.

Since the deep neural network of the baseline system was trained only on the 'Making Sense of Sounds' data set, with no external data used, and incorporates no semantic knowledge or world model, its classification must be exclusively based on the acoustic properties of the target sounds. Since these sounds appear to be rather diverse, two alternative (but not exclusive) hypotheses can be posed:

1. An unwanted and unnoticed bias in the recording situation or sound clip preparation facilitates the classification.
2. The sound-types within each of the high-level categories share some acoustic characteristics.

The first hypothesis describes a technical issue. For instance, all the sounds in the *Music* category could have been recorded with microphones of better quality and in a quieter surrounding. Thus, the baseline system would only need to use the channel characteristics to categorise a sound as music. The fact that the sounds were sourced from data bases where a diverse field of users contribute individual clips, recorded and prepared under a wide variety of circumstances, makes this hypothesis very unlikely to hold.

[4]http://cvssp.org/projects/making_sense_of_sounds/site/challenge/

The second hypothesis would entail that there is sufficient acoustic information to discriminate between the categories. Whether humans actually use this information remains unclear. It is, however, an exciting thought that these abstract categories might have some acoustic grounding even though it might only be a contributing factor in human classification, not a decisive one. Further psychological research is clearly needed here.

If the acoustic grounding would be confirmed, it would have far reaching implications. In this case, the sounds of the different high-level categories might, for instance, have different impact on humans when exposed to them over long durations [22]. If machine classifiers could reach high reliability in real-world situations, sound profiles of arbitrary locations could be compiled and set into relation to e.g. health-related demographic data at those locations.

In applied work in the 'Making Sense of Sounds' project the high-level categories are already used as the primary user-controlled filter option in custom-made hardware devices designed as tangibles interfaces for the recording and playback of sound memories [23]. The *Audio Memories* system, which encourages joint reminiscing, e.g. within a family, is to classify new sound recordings according to four of the categories (*Nature*, *Human*, *Music* and *Urban*) and to allow the user to choose them in the playback. A planned user study with target families will investigate what role the derived categories play in sound-based recall.

7. CONCLUSION

We introduced the 'Making Sense of Sounds' acoustic data set and the associated machine learning challenge, aiming at a high degree of generalisation in machine classification by making high-level human-derived categories the target. A deep learning-based baseline system performed strongly and reached an average accuracy of 80.8% on the evaluation data set.

It remains an open question whether machine and human classification share any underlying principles or even use similar acoustic features. It is also unclear whether the automated ability to classify an acoustic signal into the given categories would bring about better overall performance in more specialised tasks (e.g. through a top-down classification procedure). This, however, might be of minor importance with regard to applicability: Machine systems interacting closely with humans might simply need to have this ability for a smooth integration into human environments and it is unlikely that they have seen all relevant data in their training, forcing generalisation.

8. REFERENCES

[1] G. A. Miller, "WordNet: A lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp.

39–41, 1995.

[2] M. Sigman and G. A. Cecchi, "Global organization of the WordNet lexicon," *Proceedings of the National Academy of Sciences*, vol. 99, no. 3, pp. 1742–1747, 2002.

[3] M. Koppenberger, P. Herrera, O. Celma, and V. Tarasov, "Sound effect taxonomy management in production environments," in *Audio Engineering Society Conference: 25th International Conference: Metadata for Audio*, 2004.

[4] P. Cano, M. Koppenberger, P. Herrera, S. Le Groux, J. Ricard, and N. Wack, "Nearest-neighbor generic sound classification with a WordNet-based taxonomy," in *Audio Engineering Society Convention 116*, 2004.

[5] W. W. Gaver, "What in the world do we hear?: An ecological approach to auditory event perception," *Ecological Psychology*, vol. 5, no. 1, pp. 1–29, 1993.

[6] G. Roma, J. Janer, S. Kersten, M. Schirosa, P. Herrera, and X. Serra, "Ecological acoustics perspective for content-based retrieval of environmental sounds," *EURASIP Journal on Audio, Speech, and Music Processing. 2010: 960863.*, 2010.

[7] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[8] J. Fan, M. Thorogood, and P. Pasquier, "Automatic soundscape affect recognition using a dimensional approach," *Journal of the Audio Engineering Society*, vol. 64, no. 9, pp. 646–653, 2016.

[9] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[10] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013.

[11] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.

[12] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events*, 2017.

[13] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D. P. W. Ellis, and A. Mesaros, *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Tampere University of Technology, Tampere, Finland, 2018.

[14] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, "Bird detection in audio: A survey and a challenge," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016.

[15] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, X. Z. Fern, et al., "The 9th annual MLSP competition: New methods for acoustic classification of multiple simultaneous bird species in a noisy environment," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2013.

[16] O. Bones, T. J. Cox, and W. J. Davies, "Sound categories: Category formation and evidence-based taxonomies," *Frontiers in Psychology*, vol. 9, 2018.

[17] F. Husson, J. Josse, and J. Pages, "Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data," Tech. Rep., Applied Mathematics Department, 2010.

[18] M. J. Greenacre, *Theory and Applications of Correspondence Analysis*, Academic Press, 1984.

[19] S. Lê, J. Josse, and F. Husson, "FactoMineR: An R package for multivariate analysis," *Journal of Statistical Software*, vol. 25, no. 1, 2008.

[20] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015, pp. 1015–1018.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[22] A. V. Moudon, "Real noise from the urban environment: how ambient community noise affects health and what can be done about it," *American Journal of Preventive Medicine*, vol. 37, no. 2, pp. 167–171, 2009.

[23] T. Duel, D. M. Frohlich, C. Kroos, Y. Xu, P. J. B. Jackson, and M. D. Plumbley, "Supporting audiography: Design of a system for sentimental sound recording, classification and playback," in *HCI International: 20th International Conference on Human-Computer Interaction*, Las Vegas, Nevada, USA, 2018.